摘要

随着数据库和信息技术的迅猛发展，通过其得到的快速增长的海量数据因为得不到人们的理解而变为一座座的“数据坟墓”。作为解决这一问题的重要方法，数据挖掘引起了信息产业界的极大关注，对其相关领域技术与方法的研究已成为一个热门学科。作为数据挖掘技术的重要组成部分，关联规则挖掘技术因其广泛应用于发现大量事务商务记录中的相关关系而得到了人们特别的重视。供应链管理则有助于企业对其物流进行更好的优化配置，从而得到越来越多企业的青睐。以上几个方面的研究都在电子商务中占有重要的地位，并日益成为此领域内的研究热点。然而传统技术均存在各种各样的不足，因此迫切需要性能更好、效率更高的新方法。

本文首先引入了人工免疫系统概念，并应用其中的免疫克隆算法对多维关联规则挖掘算法及供应链管理模型求解算法进行了改进。实验证明，改进后的算法无论是在执行效率上，还是在收敛速度上，都比传统算法有了一定的提高。由于克隆算法本身具有的并行性和易操作性，使多维关联规则挖掘这类 NP 问题得以快速的解决，同时也使供应链求解算法更为快速有效。

本文分析了传统关联规则挖掘方法（Apriori 算法、基于进化的关联规则挖掘算法，基于免疫的关联规则挖掘算法）、传统供应链求解算法（基于进化算法的供应链求解算法）的优势和不足，并以此为基础，结合免疫克隆算法，做出了如下创新：

1. 关联规则挖掘是 NP 完全问题。免疫克隆算法具有比遗传算法更加优良的全局和局部寻优能力，本文给出了采用免疫克隆算法进行多维数据挖掘的步骤，并进行了仿真试验。实验证明此算法比传统算法具有更好的性能。
2. 提出了基于免疫克隆算法的供应链求解算法。在本文中，通过克隆算法来进行供应链的求解，并进行仿真试验，理论分析和试验结果表明，该算法是可行和有效的，最终取得了比较好的检测结果。本文提出的新方法为供应链求解算法的发展提供了一条新思路。

在本文的最后，对本文提出的算法与传统算法，即 Apriori 算法、基于进化的关联规则挖掘算法、基于进化的供应链求解算法做了对比，认为：传统方法与免疫克隆方法的融合是提高关联规则挖掘算法和供应链求解算法性能的新途径。

关键词：人工免疫系统　免疫克隆算法 数据挖掘　关联规则挖掘　供应链管理

Abstract

Along with the rapid development in database and technology technologies, plentiful data we got from it has become data tomb because it can't be understood by us. As the resolution of this problem, data mining becomes the focus of the IT world.As the important part of data mining, association-rule mining attracts people heavily because it can be used to find interesting relations from lots of data.At the same time,supply chain can help companies to disposing their resources better.These technologies are used widely in E-commerce.

First, we introduce the conception of artificial immune system (AIS) in this article. Then we modify the association-rule mining algorithm and supply chain solution algorithm by the immune clonal algorithm in artificial immune system. The experiments show that if we use the algorithms provided in this article, we get better execution efficiency and faster convergence velocity than traditional algorithms .And we can get better supply chain solution by the methods provided in this article.

Then, we analyzed the advantages and shortcomings of traditional association-rule mining algorithms and supply chain solution algorithm.Based on these results and immune clone algorithm, we provide some methods as follows:

1. Based on the fact that immune clonal algorithm has better ability both at overall and partial searching, we introduce the multi-division association-rule mining algorithm based on the immune clone algorithm (AMBC). The experiments show that it has better performance than traditional algorithms.
2. We introduce the supply chain solution algorithm based on the immune clonal algorithm (SCBC). We find it has good performance both at convergence velocity and searching ability though the experiments.

At last, we provide two comparisons, one is between AMBC and traditional algorithms (Apriori algorithm, AMBE&AMBC), the other is between SCBC and SCBE. We draw such a conclusion: it is a novel significant way that we combine the immune clonal algorithm with association-rule mining and supply chain solution.

Key words: artificial immune system, immune clonal algorithm, data mining, association-rule mining, supply chain management.

创新性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：[signature] 日期 2015年2月28日

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本学位论文属于保密在___年解密后适用本授权书。

本人签名：[signature] 日期 2015年2月28日

导师签名：[signature] 刘方 日期 _____

第一章 绪 论

1.1 研究背景和意义

1.1.1 数据挖掘

近年来，数据库技术的飞速发展以及数据库管理系统的广泛应用，使人们积累了大量的数据。数据的丰富带来了对强有力的数据分析工具的需求，大量的数据被描述成“数据丰富，但信息贫乏”。快速增长的海量数据被收集和存放在大型数据库中，没有强有力的工具，要理解它们已经远远超出了人的能力。为了使决策者从海量数据中提取有价值的信息，数据挖掘这一概念便应运而生。数据挖掘的目的就是从巨量的数据集合中提取隐含的、先前未知的、对决策有潜在价值的规则[1]。

关联规则挖掘发现大量数据中项集之间有趣的关联或相关关系。随着收集和存储的数据不停地增多，许多业界人士对于从他们收集和存储的数据中挖掘关联规则越来越感兴趣。从大量商务事务记录中发现有趣的关联关系，可以帮助许多商务决策者制定有效的决策，如分类设计、交叉购物和贱卖分析等。

1.1.2 供应链管理

物流是指物质实体从供应者向需求者的物理移动，它由一系列创造时间价值和空间价值的经济活动组成，包括运输、保管、照送、包装、装卸、流通加工及物流信息处理等多项基本活动，是这些活动的统一。

长期以来，物流领域的技术含量低于社会平均水平，属于劳动密集型行业．劳动生产率低，运营成本高，严重阻碍了生产与流通领域的顺畅发展。

其实物流领域是最应该也最可能实施现代化技术的领域。由于其在经济发展中的重要作用而产生的推动力及其现状所反映出的巨大发展潜力，物流业一定会迅猛地发展起来，其发展的途径与方向一定是信息化与三流(物流、信息流、资金流)统一。而实现这种统一的基础就是供应链及其管理。

1.2 目前研究状况及发展方向

1.2.1 关联规则挖掘

关联规则的挖掘最早是由 Agrawal，Imielinski 和 Swami 提出[2]。Apriori 算法是由 Agrawal 和 Srikant 提出的[3]。使用剪枝方法的算法变形由 Mannila,Toivinen 和 Verkamo 独立开发[4]。Agrawal 和 Shafer 提出了统计分布 CD 和数据分布 DD 的并

行的 apriori 算法[5]。Toivonen 讨论了选样方法[6]，Brin，Motwani，Ullman 和 *Tsur* 提出了动态项集计数方法[7]。关于关联规则挖掘的许多改进的方法中，*Han* 和 *Kumar* 等人研究了关联规则的尺度化的并行数据挖掘算法[8]，Agrawal 和 Srikant 提出了序列模式挖掘[9]，Koperski 和 Han 研究了空间关联规则的挖掘[10]，Lu,Han 和 Feng 给出了事务间的关联规则的挖掘[11]，最大模式的挖掘是由 Bayardo 给出的[12]，大闭项集的挖掘是由 Pasquier,Bastide,Taouil 和 Lakhal 提出的[13]，大项集的深度优先算法由 Beyer 和 Ramakrishnan 提出[14]，挖掘大模式而不产生候选项集的方法由 Han,Pei 和 Yin 提出[15]。

1.2.2 供应链模型应用现状

由于采用供应链管理可以使企业在多方面获得实在与潜在的收益，优化运行状况，所以对它的研究与实践越来越受到重视。从而产生了多个适合于我国国情的各个系统的供应链模型。李群明等研究了基于 COBRA 的多 Agent 供应链管理系统[16]，王春喜等研究了面向网络制造的知识供应链建模[17]，赵耀华研究了以核心业务为中心的敏捷供应链[18]，袁磊等研究了数据仓库的信息供应链模型[19]，曹文彬等研究了流程企业供应链管理的建模问题[20]，邵晓峰等研究了面向大规模定制的供应链模型[21]，韩向东等研究了基于标准过程的供应链管理模型[22]，卢震等研究了供应销售条件下的供应链模型与决策应用[23]。为了改进供应链模型的各项运行参数，也出现了许多新方法，其中黄小原等研究了交互式进化规划在供应链方向的应用[24]，曹杰等研究了供应链联合优化数学模型及求解的混合算法[25]，卢震等在服务销售系统供应链模型设计与应用中应用了遗传算法[26]。

1．3 论文研究的意义和所做的工作

最典型的关联规则挖掘算法是 Apriori 算法，但 Apriori 算法执行效率较低、要产生大量的候选项集、并且要频繁的重复扫描数据库。基于 Apriori 算法的这些缺点，提出了许多改进型的 Apriori 算法，其中有基于散列技术、事务压缩技术、划分技术、选样技术和动态项集计数的改进型的 Apriori 算法，另外还提出了不产生候选挖掘频繁项集的使用频繁模式增长策略的关联规则挖掘算法。虽然每种算法均对原始 Apriori 算法的一些缺点进行了改进，但它们均不同程度的出现收敛速度慢，所挖掘出的关联规则准确率不高等问题。

供应链管理可以在多方面使企业受益：如降低库存量，增加库存周转；减少纸上作业，加快存周转；减少纸上作业，加快支付过程；缩短产品周期；降低运输成本；消去多余职责，优化通信结构，减少错误等。因此各种各样的模型如雨后春笋被设计出来以满足众多企业的需要。这些供应链模型一般将实际的生产销

售问题化解为一个混合整数规划问题，如果找不到合适的求解规划的方法，则可能会大大降低此供应链模型的实用性及可操作性。

进化算法(即 EA)是基于种群进化繁殖思想发展起来的一类随机搜索技术。它们是模拟由个体组成群体的集体学习过程。其中每个个体表示给定问题搜索空间中的一点。进化算法从任一初始的群体出发，通过随机选择、交异和交叉过程,使群体进化到搜索空间中越来越好的区域。进化算法具有优化过程快、易于并行、扩展性强等优点，并且它易于同其他技术混合、可以较快的解决复杂的问题。而当今关联规则挖掘和混合整数模型求解均面临大规模和高维的困难，进化算法由于具有以上优点，所以在关联规则挖掘和求解混合整数模型时采用进化算法能在一定程度上解决其面临的问题。

克隆选择算法是一种新的人工免疫系统方法。它不仅具有进化算法的所有优点，而且兼顾全局搜索与局部搜索，弥补了进化算法较少关注种群间的协作这一缺点，从而具有更好的种群多样性，并且克隆选择算法具有良好的记忆效应。所以无论在理论证明上还是实践结果上，克隆选择算法都优于与其对应的遗传算法。而其中包含的多克隆算法由于加入了交叉的操作，从而具有更为优秀的性能。所以将多克隆选择算法应用于关联规则挖掘和供应链模型求解是可行的，也是有意义的。

在课题研究中，首先研究了数据挖掘及其挖掘技术的体系，然后对关联规则挖掘进行了深入的研究，提出了将多克隆选择算法应用于多维关联规则挖掘的改进算法，通过对算法的实现，并通过对某销售记录进行仿真试验，结果表明改进后的算法具有更快的收敛速度和更强的搜索能力。另一方面，通过对供应链管理概念及模型的深入研究，提出了将克隆选择算法应用于服务销售系统供应链的新算法，通过仿真试验证明本文算法具有更强的搜索能力及更快的收敛速度。

1．4 本文内容安排

全文共分为六章,分别如下：

第一章　绪论　介绍了数据挖掘、关联规则挖掘及供应链管理的发展概况，新算法研究的意义和方向以及本文所完成的主要工作；

第二章　人工免疫系统　介绍了人工免疫系统的基本概念及研究背景，同时着重介绍了其中的免疫算法和克隆算法，并对人工免疫系统的应用进行了简单描述。

第三章　数据挖掘与关联规则挖掘　讨论了数据挖掘技术的背景、任务、分类及计算智能方法在这一领域的应用现状；同时，介绍了关联规则挖掘的相关概念，了解了多层关联规则和多维关联规则的基本定义；最后，我们还提出了挖掘

关联规则的几种常用算法，以及如何衡量关联规则的价值方法。

第四章　基于多克隆选择算法的关联规则挖掘算法　提出了将免疫克隆算法应用于多维关联规则挖掘的算法。该算法既克服了传统 Apriori 算法挖掘速度慢、计算过程复杂、不能并行化处理等问题，也克服了遗传算法及免疫算法忽略局部搜索、易早熟的问题。它能够使聚类结果收敛到全局最优，实验证明，相比遗传算法和免疫算法，有效克服了早熟问题、保持了解的多样性，具有更快的收敛速度和更强的搜索能力。

第五章　供应链管理　本章首先介绍了供应链管理的概念、分类及其所采用的相关技术。作为供应链管理的重要发展方向，协作管理及其发展现状也在本章中进行了介绍。在本章最后的篇幅中，简述了供应链的构造流程。

第六章　基于克隆算法的供应链求解算法　针对传统进化算法运行速度比较慢且容易早熟的缺点，主要分析和研究了克隆算法在供应链求解中如何应用等问题。通过充分利用克隆算法优良的全局及局部搜索能力，从而使算法具有更好的种群多样性和更快的收敛速度。通过仿真实验，证明了本文算法在求解供应链问题时的优良性能。

第七章　总结与展望

第二章 人工免疫系统

近年来，生物免疫系统(immune system) 已成为一个新兴的生物信息研究课题[27]。免疫系统是由器官、细胞和分子组成的一个复杂系统，它是除神经系统外，机体能特异地识别“自己/ 非己”刺激，对之作出精确应答，并保留记忆的功能系统。研究表明，免疫系统具有多种功能[28]：如模式识别、学习、记忆获取、多样性、容错及分布式检测等。本文主要研究其中的免疫计算方法[29, 30]。

2.1 免疫系统及其人工免疫系统简介

2.1.1 免疫系统概述

免疫系统对侵入机体的非己成分(如细胞、病毒和各种病原体) 以及发生了突变的自身细胞(如癌细胞) 具有精确识别、适度应答和有效排除的能力[31]。免疫系统的功能主要由免疫细胞完成，T 细胞和 B 细胞是淋巴细胞的 2 种主要类型。B 细胞由骨髓产生,并通过分泌抗体(antibody) 与抗原(antigen) 相结合，实现抗原的清除。T 细胞由胸腺产生，在免疫反应过程中能刺激和抑制 B 细胞的增殖和分化，对免疫调节起着重要的作用[32]。免疫系统可分为先天性免疫系统和适应性免疫系统。先天性免疫系统是生物在种系发育和进化过程中逐渐建立起来的一系列天然防御功能，其特点是与生俱有，能传给下一代，无特异性，对各种病原体都有一定的防御功能。适应性免疫系统在“初次响应”之后，可记住病原体，当再次遇到时会产生“再次响应”，迅速而有效地消除病原体。

2.1.2 免疫系统的信息处理特性

免疫系统具有以下特性：

1) 多样性　免疫系统的抗体库具有多样性,能及时有效地消除不同的入侵抗原；

2) 容错性　在分类和响应中的一些小错误不会酿成大错；

3) 分布自律性　由许多局部相互作用的基本单元组成来提供全局的保护，没有集中控制；

4) 动态稳定性　免疫系统要消除不断变化的外来抗原入侵并保持系统的稳定；

5) 自适应鲁棒性　它的强大的学习能力使之成为随环境改变而不断完善的一个自适应鲁棒进化系统。

2.1.3 人工免疫系统的研究背景

人工免疫系统(Artificial immune system ，简称 AIS) 的研究起步相对较晚，但近两年来发展迅速。1996 年在日本召开了第 1 次“基于免疫的系统”国际学术讨论会(IMBS96)[33]，1997 年和 1998 年的 IEEE Systems ,Man and Cybernetics 国际会议对 AIS 组织了专题讨论。1998 年出版了论文集《人工免疫系统及其应用》[27]，提出多种方法定义形式上的受免疫概念启发的各种理论研究和工程应用：人工免疫系统、基于免疫的系统、人工免疫网络等。2002 年 9 月,在英国的 Kent 大学召开第一届人工免疫系统国际大会,会议的目的是通过研究不同的免疫学机制和它们在信息处理和工程设计中的关系,来加强 AIS 的研究。

2．2 人工免疫系统中的主要方法

人工免疫系统的方法主要有基于免疫网络学说的人工免疫网络模型[34]、基于免疫特异性的否定选择算法[30]、免疫进化算法[29]、免疫克隆算法[35]。本文应用了其中的免疫进化算法和免疫克隆算法。

2.2.1 免疫进化算法

近 20 年中，进化算法作为一种随机搜索的优化方法，具有通用性、便于并行处理等优点，得到了广泛应用。然而在实际应用中也存在一些需改进之处：早熟现象，算法的稳定性与收敛速度的关系等。免疫系统是一个随环境改变而不断进化的系统。将进化与免疫结合起来考虑,能得到更有效的优化算法。 这里将介绍 2 种免疫进化算法：①有效利用先验知识的免疫算法； ②具有种群多样性的免疫遗传算法。

1.利用先验知识的免疫算法

进化算法本身是一种不依赖问题的通用方法，而每一个实际问题都会有自身一些基本的、显而易见的特征信息或知识，进化算法在求解问题时无法利用这些先验知识，使问题求解不具有针对性。文献[29]在标准的遗传算法中引入免疫概念，有效地利用问题的先验知识，提出了免疫算法，在保留原算法的优良特性的前提下，引入了一个新的算子——免疫算子(immune operator) ，有选择、有目的地利用待求问题中的一些特征信息或知识，提取“疫苗”来抑制其优化过程中出现的退化现象。在进化选择过程中，通过“接种疫苗”和“免疫选择”来指导搜索过程，获得更好的优化性能。与通用遗传算法相比，免疫算法较好地解决了已有算法中出现的退化现象，且使收敛速度有显著提高。

2.利用抗体多样性的免疫遗传算法

在进化算法中，如果根据适应度函数选出的双亲基因非常接近，那么所产生的后代相对双亲也必然比较接近，这样所期待的改善就比较小，基因模式的单一性不仅减慢进化历程，而且可能导致进化停滞，过早收敛于局部最优点。

抗体多样性是免疫系统的一个重要特性，在免疫调节中，那些有高抗原亲和力并且浓度较低的抗体会受到促进，而低抗原亲和力及浓度较高的抗体将会受到抑制，以此保证抗体的多样性。将这一概念应用到标准的进化算法中，在群体中个体浓度越小，选择概率越大，个体浓度越大，则选择概率越小，可以在保留高适应度个体的同时，进一步确保个体多样性，改善早熟现象。

2.2.2 免疫克隆选择算法

1.克隆选择算法简介

进化算法是在繁殖、变异、竞争、选择等生物模型的基础上形成的，它对生物繁殖机理的模仿集中在有性繁殖上[36][37]，有性繁殖能实现代与代之间的信息交换，强调新信息的产生。然而，当抗原侵入生物机体时，其免疫系统在机体内选择出能识别和消灭响应抗原的抗体，这一过程主要借助克隆（无性繁殖）使机体内的抗体激活、分化和增殖，以增加其数量，进一步进行免疫应答并最终清除抗原[38][39]，因为抗原由载体和半抗原组成，是免疫应答的始动因子；抗体是免疫应答的重要产物。而抗体由两条重链和轻链组成，包括恒定区和可变区。抗体与抗原的结合正是在高可变区，该区也是免疫系统多样性的物质基础。抗体除了有识别抗原分子的抗原决定簇外，还可以和其它抗体分子结合的对位，因此抗体具有识别抗原又能被其它抗体识别的双重性。

在人工智能计算中[40]，抗原、抗体一般分别对应于求解问题及其约束条件和优化解。因此，抗原与抗体的亲和度（即匹配程度）描述解和问题的适应程度，而抗体与抗体间的亲和度反映了不同解在解空间中的距离。亲和力就是匹配程度。由于抗体的双重性，因此这里的亲合度也包括两种，即抗体与抗体间的亲合度，以及抗体抗原间的亲合度。因此，定义作用于抗体与抗原的亲和力成熟算子为亲合力成熟算子 I；定义作用于抗体与抗体的亲和力成熟算子为亲合力成熟算子 II。

亲合力成熟算子 I 作用于抗体和抗原，使其亲合度增强，就意味着在解空间，使侯选解更能满足问题的需要。一般采用抗体（解）的目标函数（抗原）值或其函数值作为抗体与抗原间的亲和度。

对于抗体群 $A(k)$，显然存在 M_A，有：

$$A(k+1)=\Omega^A(A(k))=A(k)+M_A$$

使得亲和度增加。

如果抗体群 $A(k)$采用二进制编码，M_A 可以是从父代继承的变化规律。而对于

实数编码的抗体群，为了增加抗体与其的亲和度最常用的方法是爬山法，即根据目标函数$f(X)$的一阶导数确定梯度最大的方向，然后沿此方向迁移。即：

$$M_A = D\frac{d(f(X))}{dX}\Big|_{X=A(k)}$$

D是步长。实践中，如果不能求得$f(X)$的导数，则取上一次最优的变化。

总之，亲合力成熟算子I是为了提高算法的局部搜索能力，因此，经过适当改进的传统局部搜索算法都可以成为该算子。通过亲合力成熟算子I，一个完整的人工免疫系统就有机地将全局搜索和局部搜索结合起来，有利于问题的求解。

亲合力成熟算子 II 作用于抗体与抗体，使其间亲合度增加，意味着彼此距离的增加，即增加了种群的多样性，也就是抑制无效抗体的产生。祛除抗体群 $A(k)$ 中和彼此距离近且与抗原亲合度小的抗体，即：

$$C(A(k)) = A'(k) = \{a_i(k) \mid a_i(k) \in A(k) \ \ and \ \ d_{ij} > \sigma(k) \ \ and \ f(a_i(k)) > \eta\}$$

其中，d_{ij}是抗体间的距离(亲合度)，$\sigma(k)$和η是相应的阈值。如果抗体群$A(k)$采用二进制编码，d_{ij}一般取海明距离，对于实数编码，d_{ij}一般取欧氏距离

2.抗体克隆机理

抗体克隆选择学说认为，当抗原侵入机体时，克隆选择机制能在机体内选择出能识别和消灭相应抗原的免疫细胞，使之激活、分化和增殖，进行免疫应答以最终清除抗原，这就是克隆选择[38]。在这一过程中，克隆这一无性繁殖过程中父代与子代间只有信息的简单复制，而没有不同信息的交流，无法促使抗体种群进化。为了在人工智能中借鉴这一机理，需要提出新的克隆算子。

克隆选择是由亲合度诱导的抗体随机映射，抗体群的状态转移情况可以表示成如下的随机过程：

$$C_S: \ A(k) \xrightarrow{clone} A'(k) \xrightarrow{mutation} A''(k) \xrightarrow{compress} A(k+1)$$

同前面介绍的一样，抗原、抗体、抗原和抗体之间的亲和度分别对应于优化问题的目标函数和各种约束条件、优化解、解与目标函数的匹配程度。那么克隆算子就是依据抗体与抗原的亲合度函数$f(*)$，将解空间中的一个点$a_i(k) \in A(k)$分裂成了q_i个相同的点$a_i'(k) \in A'(k)$，经过克隆变异和克隆选择变换后获得新的抗体群。因此，克隆算子实际上包括了三个步骤，即克隆、克隆变异和克隆选择。对于二进制编码，抗体$a \in B^l$ ，其中，$B^l = \{0,1\}^l$代表所有长度为l的二进制串组成的集合，抗体群$A = \{a_1, a_2 \cdots a_n\}$为抗体$a$的$n$元组。

3.克隆算子

克隆算法是依靠编码来实现与问题本身无关的搜索，一般可采用二进制编码和十进制编码两种方法，本文采用十进制编码方法。我们可以把相关属性用相应

的整数值代替，从而产生种群数目为 n 的抗原群体。

抗体-抗原亲和力函数 f 一般是目标函数 φ 的函数，抗体-抗原亲和力定义为：

$$D_{ij} = \| A_i - A_j \| \quad i,j = 1,2,\ldots,n$$

D_{ij} 为任意范数，对二进制编码一般取海明距离，而十进制编码多取为欧氏距离。记 $D = (D_{ij})_{n*n}$ i,j=1,2,…,n 为抗体-抗体亲和力矩阵。D 是一对陈矩阵，反映了种群的多样性。

克隆操作 T_c^C：定义

$$T_c^C(\overline{A}(k)) = [T_c^C(A_1(k)) \ldots T_c^C(A_n(k))]^T$$

其中：$T_c^C(A_i(k)) = I_i * A_i(k)$ i=1,2,…n，I_i 为 q_i 为行向量，称抗体 A_i 的 q_i 克隆。

$$q_i = g(N_c, f(A_i(k)))$$

一般取：

$$q_i = Int\left(N_c * \frac{f(A_i(k))}{\sum_{j=1}^{n} f(A_j(k))} \right) \qquad i = 1,2,\ldots,n$$

N_c>n 是与克隆规模有关的设定值：$Int(*)$ 为上取整函数，$Int(x)$ 表示大于 x 的最小整数。由此可见，对单一抗体而言，其克隆规模是根据抗体-抗原亲和度自适应调整的。克隆过后，种群变为

$$\overline{A}'(k) = \left\{ \overline{A}(k), \overline{A_1}'(k), \overline{A_2}'(k), \ldots, \overline{A_n}'(k) \right\}$$

其中：

$$\overline{A_i}'(k) = \left\{ A_{i1}(k), A_{i2}(k), \ldots, A_{i(q_i-1)}(k) \right\} A_{ij}(k) = A_i(k) \quad j = 1,2,\ldots,q_i - 1$$

免疫基因操作 T_g^C：免疫基因操作主要包括交叉和变异。仅是用变异的克隆算子为单克隆算子；交叉和变异都采用的为多克隆算子。

依据概率 p_m^i 对克隆后的群体进行变异操作，$\overline{A}''(k) = T_g^C(\overline{A}'(k))$，为了保留抗体原始种群的信息，便以算子并不作用到 $\overline{A}(k) \in \overline{A}'(k)$，即：

$$P_g(A_{ij}(k) \longrightarrow A_{ij}'(k)) = \begin{cases} (P_m^i)^{H(A_{ij}(k), A_{ij}'(k))} (1 - P_m^i)^{1 - H(A_{ij}(k), A_{ij}'(k))} & A_{ij}(k) \in \overline{A_i'(k)} \\ 0 & A_{ij}(k) \notin \overline{A_i'(k)} \end{cases}$$

免疫选择操作 T_s^C：$\forall i = 1,2,\ldots,n$，存在变异后抗体 $B = \left\{ A_{ij}'(k) \mid \max f(A_{ij}') j = 1,2,\ldots,q_i - 1 \right\}$，则 B 取代 $A_i(k) \in \overline{A}(k)$ 的概率为：

$$p_s^k(A_i \longrightarrow B)=\begin{cases}1 & f(A_i(k))<f(B)\\ \exp(-\dfrac{f(A_i(k))-f(B)}{a}) & f(A_i(k))\geq f(B)\text{且}A_i(k)\text{不是目前种群的最优个体}\\ 0 & f(A_i(k))\geq f(B)\text{且}A_i(k)\text{是目前种群的最优个体}\end{cases}$$

a>0 是一个与抗体种群多样性有关的值，一般的多样性越好，a 取值越大，反之越小。

克隆死亡操作 T_d^C：克隆选择后获得相应的新抗体群为：

$$\overline{A}(k+1)=\{A_1(k+1),A_2(k+1),\ldots,A_i'(k+1),\ldots,A_n(k+1)\}$$

其中，

$$A_i'(k+1)=A_j(k+1)\in\overline{A}(k+1)$$

$$i\neq j\text{且}f(A_i'(k+1)=f(A_j(k+1)=\max f(\overline{A}(k+1))$$

那么，以概率 p_d 任意死亡 $A_i'(k+1)$与$A_j(k+1)$ 中的一个。死亡策略既可以是随机产生一个新抗体代替 $A_i'(k+1)$或$A_j(k+1)$，也可以是采用变异或交叉策略重新生成新抗体代替 $A_i'(k+1)$或$A_j(k+1)$。

克隆算子作用后获得相应的新抗体群为 $\overline{A}(k+1)=\{A_1(k+1),A_2(k+1),\ldots,A_n(k+1)\}$，返回第一步循环执行克隆算法。

4.多克隆选择算法

仅采用克隆算子的克隆选择算法称为简单克隆选择算法，采用多克隆算子的克隆选择算法称为多克隆选择算法，算法流程如下所示：

多克隆选择算法（Poly-Clonal Selection Algorithm）

step1　k=0，初始化抗体群落 $\overline{A}(0)$，设定算法参数，计算初始种群的亲和度；

step2　依据亲和度和设定的抗体克隆规模，进行多克隆算子操作，获得新的抗体群落 $\overline{A}(k)$；

step3　k=k+1；若满足停止条件，终止计算；否则，回到 step2。

实际应用中一般采用限定迭代此书或在连续几次（如 t 次）迭代中记忆单元的最好解都无法完善，以及二者的混合形式作为终止条件。

5.克隆选择算法的优点

克隆算法与进化算法都是群体搜索策略，并且强调群体中个体间的信息交换，因此具有很多相似之处。无论是算法结构、本质上具有的固有并行性和搜索变化的随机性、与其它只能策略结合的固有优势、主要算子的计算方法，都存在很多的共性。

但克隆选择算法不是进化算法的简单改进，而是新的人工免疫系统方法。首先，克隆算法的基本思想来源于免疫系统而非自然进化；其次，在算法的实现上，进化算法更多地强调全局搜索，而忽视局部搜索，但克隆选择算法两者兼顾，有

更好的种群多样性；再次，克隆选择算法弥补了进化算法较少关注种群间的协作这一缺点，提出了抗体-抗体亲和度概念；最后，克隆算法中变异是主要算子，交叉是次要算子，与进化算法相反，实践中也证明了克隆选择算法的性能强于相应的遗传算法；另外，克隆算子具有记忆性能，因此本身就能保证算法以概率 1 收敛到最优解，而遗传算法则没有这一特性。

2.3 人工免疫系统的应用现状

从研究方法及使用的免疫概念来看，主要的应用研究可分为以下几条主线：

1) 构造动态的人工免疫网络模型，从结构上模拟免疫系统的动态特性，使系统具有高度的自治性和学习能力，为人工智能的研究开辟了一条新的思路。

2) 从功能上模拟免疫系统，使系统具有识别异己并对外来入侵及时反应。这些研究在计算机及其网络安全领域有着广阔的应用前景。

3) 与进化算法相结合，从演化的角度进行免疫模拟。在进化算法中引入免疫概念，改进候选解的种群多样性，提高进化算法收敛到全局最优解的能力，可得到更有效的优化算法。

4) 其他免疫概念的应用，如免疫反应的特异性概念对故障检测有一定的指导意义；免疫调节机制对控制的启发等。

2.3.1 人工智能

学习是人工智能研究的一个重要方面，如果一个系统的行为是由自身的当前输入和过去的经验而不是设计者的输入和经验来决定，并且具有从经验中获得知识并利用知识来解决新的未曾遇见难题的能力，则认为该系统具有学习能力。免疫化过程(通过疫苗) 是一种免疫学习典型的例子，另外免疫系统的分布式特性，特别适用动态环境中的信息处理过程，这为人工智能研究注入了新的生命力。

1.机器学习

文献[41]提出了基于免疫网络理论的人工免疫系统(artifical immune system ,AIS)，并进行了机器学习的研究，AIS 提供噪声耐受、无监督学习，不需要学习知识的精确表达及反例。这种系统聚合了学习分类系统、神经网络、机器推理和基于案例的检索各类方法的优点。AIS 的运行包括一个根目标，一个细胞网络、一组示教数据和一组测试数据。网络中的每个细胞处理一个模式识别基本单元，这一单元是由模拟自然免疫系统中的抗体形成的基因机制。这就允许使用复杂的词汇并增强了模式匹配基本单元的多样性。AIS 成功地进行了在 DNA 序列中促进剂的识别。

2.数据挖掘

在人工智能的研究中，知识的自动获取将是一个关键，数据挖掘技术是解决这一关键的主要方案。文献[42]比较了人工免疫网络、聚类分析和神经网络三种方法在数据挖掘中的应用和各自特点，指出应用人工免疫系统进行数据挖掘，可对训练数据进行建模，对输入空间的大区域有泛化能力，并能对得到的进化网络提供更好的解释，获取更多的有用信息。

2.3.2 计算机安全领域

随着计算机系统及其互联网的高速发展，计算机网络安全成为日益突出的问题，而防御异常入网络安全成为日益突出的问题，而防御异常入侵、防范病毒等都可以从生物免疫机制中获得不少启发。文献[43]将人工免疫网络的分布性、鲁棒性、动态性、多样性和自适应性应用到计算机网络的安全领域。

2.3.3 智能控制

实际的工业控制对象具有非线性、不确定性、参数分布性和时变性等复杂特性，传统的控制方法无法获得更好的控制效果。而将免疫机制引入控制领域，对解决复杂的动态自适应控制难题提供了崭新的思路。

2.3.4 其他工程领域

在人工神经网络优化设计和智能建筑等其他智能控制领域也有免疫思想的渗透。文献[44]将免疫进化算法用于人工神经网络的设计，提出了基于免疫调节的共生进化网络设计方法。

2.4 本章小结

人工免疫系统是近年来新产生的一门学科，发展十分迅速，它借鉴生物免疫系统建立起来的，它具有自然免疫系统所具有的多样性、容错性、分布自律性、动态稳定性、自适应鲁棒性等优点。

在人工免疫系统体系中，人们提出了一些新颖的免疫算法，其中有免疫算法、克隆算法等，这些算法均借鉴了自然群体的有关特性，具有良好的执行性能。

随着相关理论及方法的逐步完善，人工免疫系统已被广泛应用于机器学习、数据挖掘、智能控制等许多领域。

第三章 数据挖掘与关联规则挖掘

3.1 数据挖掘

3.1.1 数据挖掘技术的起源

数据挖掘是从大量数据中识别出有效的、新颖的、潜在有用的，以及最终可理解的知识和模式的高级操作过程[1]。而数据挖掘技术是应用需求推动下多种学科融合的结果，这些学科领域包括数据库技术、统计分析、机器学习、智能计算、模式识别、神经网络、数据可视化、信息检索、图像数据库与信号处理，及空间数据分析等。但是与数据挖掘技术发展密切相关的是以下二个领域技术的飞速发展。

首先是数据库技术，随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，不缺数据缺知识的矛盾日益突出，如何理解已有的历史数据并用以预测未来的行为，如何从这些海量数据中发现知识，为决策者提供重要的决策，从而获取更大的经济效益和社会效益。从数据库的角度看，数据挖掘技术的重点是设计面向大规模数据的高效率的、适应性强的算法。目前的数据库技术不仅可以实现数据的高效查询、统计等功能，而且还可以对大规模数据集进行选择或投影操作以及分布式数据库技术等这些都为数据挖掘技术中处理海量数据提供了有效的支持。即为数据挖掘技术的应用制作了一个很好的平台。

其次是人工智能领域的一个重要分支—机器学习的研究也取得了很大进展。自从 50 年代开始机器学习的研究以来，先后经历了神经模型和决策理论、概念符号获取及知识加强、论域专用学习三个阶段，根根据人类学习的不同模式人们提出了很多机器学习方法，如：实例学习、观察和发现学习、神经网络和遗传算法等等。其中某些常用且较成熟的算法已被人们运用于实际的应用系统及智能计算机的设计和实现中。数据挖掘中的许多方法就来源于机器学习[45]。

因此，一个成熟的数据挖掘系统应该有机地结合数据库技术和人工智能技术。因为数据库技术提供了操作大规模数据的手段，同时为存储和管理发现的知识提供了良好的平台，人工智能技术则构成了学习过程的核心。但直接把数据库技术和人工智能算法合并在一起并不能满足我们的要求。在应用需求的推动下，数据挖掘技术不断发展，同时结合其它领域的研究成果，从而提高挖掘的实时性、随机性、动态性，进而开发和实现智能数据挖掘。

3.1.2 数据挖掘的任务和六种模式

数据挖掘的任务是从数据中发现模式。模式是一个用语言 L 来表示的一个表达式 E，它可用来描述数据集 F 中数据的特性，E 所描述的数据是集合 F 的一个子集 FE。E 作为一个模式要求它比列举数据子集 FE 中所有元素的描述方法简单。例如，“如果成绩在 81～90 之间，则成绩优良”可称为一个模式，而“如果成绩为 81、82、83、84、85、86、87、88、89 或 90，则成绩优良”就不能称之为一个模式。

模式有很多种，按功能可分有两大类：预测型(*Predictive*)模式和描述型(*Descriptive*)模式。

预测型模式是可以根据数据项的值精确确定某种结果的模式。挖掘预测型模式所使用的数据也都是可以明确知道结果的。例如，根据各种动物的资料，可以建立这样的模式：凡是胎生的动物都是哺乳类动物。当有新的动物资料时，就可以根据这个模式判别此动物是否是哺乳动物。

描述型模式是对数据中存在的规则做一种描述，或者根据数据的相似性把数据分组。描述型模式不能直接用于预测。例如，在地球上，70%的表面被水覆盖，30%是土地。

在实际应用中，往往根据模式的实际作用细分为以下 6 种：

1. 分类模式

分类模式是一个分类函数(分类器)，能够把数据集中的数据项映射到某个给定的类上。分类模式往往表现为一棵分类树，根据数据的值从树根开始搜索，沿着数据满足的分支往上走，走到树叶就能确定类别。

2. 回归模式

回归模式的函数定义与分类模式相似，它们的差别在于分类模式的预测值是离散的，回归模式的预测值是连续的。如给出某种动物的特征，可以用分类模式判定这种动物是哺乳动物还是鸟类；给出某个人的教育情况、工作经验，可以用回归模式判定这个人的年工资在哪个范围内，是在 6000 元以下，还是在 6000 元到 1 万元之间，还是在 1 万元以上。

3. 时间序列模式

时间序列模式根据数据随时间变化的趋势预测将来的值。这里要考虑到时间的特殊性质，像一些周期性的时间定义如星期、月、季节、年等，不同的日子如节假日可能造成的影响，日期本身的计算方法，还有一些需要特殊考虑的地方如时间前后的相关性(过去的事情对将来有多大的影响力)等。只有充分考虑时间因素，利用现有数据随时间变化的一系列的值，才能更好地预测将来的值。

4. 聚类模式

聚类模式把数据划分到不同的组中，组之间的差别尽可能大，组内的差别尽可能小。与分类模式不同，进行聚类前并不知道将要划分成几个组和什么样的组，也不知道根据哪一(几)个数据项来定义组。一般来说，业务知识丰富的人应该可以理解这些组的含义，如果产生的模式无法理解或不可用，则该模式可能是无意义的，需要回到上阶段重新组织数据。

5. 关联模式

关联模式是数据项之间的关联规则。关联规则是如下形式的一种规则："在无力偿还贷款的人当中，60%的人的月收入在3000元以下。"

6. 序列模式

序列模式与关联模式相仿，而把数据之间的关联性与时间联系起来。为了发现序列模式，不仅需要知道事件是否发生，而且需要确定事件发生的时间。例如，在购买彩电的人们当中，60%的人会在3个月内购买影碟机。

在解决实际问题时，经常要同时使用多种模式。分类模式和回归模式是使用最普遍的模式。分类模式、回归模式、时间序列模式也被认为是受监督知识，因为在建立模式前数据的结果是已知的，可以直接用来检测模式的准确性，模式的产生是在受监督的情况下进行的。一般在建立这些模式时，使用一部分数据作为样本，用另一部分数据来检验、校正模式。聚类模式、关联模式、序列模式则是非监督知识，因为在模式建立前结果是未知的，模式的产生不受任何监督。

3.1.3 数据挖掘技术的分类

现有的数据挖掘技术——预测模型化、聚类、数据归纳、依赖模型化以及发现变化和偏差，下面分别作以简要介绍：

1. 预测模型化 是在一个数据库中基于一些字段预测另一些字段。如果被预测的字段是数值型的，则预测就是一个回归问题；如果字段值是分类的形式，该问题就是一个分类问题。分类和回归技术有许多的演变，都可用于预测问题。预测的目标是利用历史数据自动推导出给定数据的推广描述，从而对未来数据进行预测。分类输出是离散的类别值，回归的输出是连续数值。一般地，这样的问题总是在给定的训练数据、一些字段、和一组问题的领域知识的表示中决定被预测变量最可能的取值。分类的基本目标是预测一个分类变量的最可能的状态。分类器的构造方法主要有统计学方法（贝叶斯方法和非参数法）、机器学习方法（决策树、决策表和产生式规则）以及神经网络方法（BP算法和粗集等方法）等。一般分类的效果与数据的特点有关，有时分类会导致数据的噪声较大、有缺省值或数据分布稀疏或字段间相关性太强等问题。所以目前还没有一个公认的适合各种类

型的数据的分类算法。

2. 聚类 把一组个体按照相似性归成若干类别，其目的是使属于同一类的个体之间的距离尽可能小，而不同类别的个体之间的距离尽可能的大。聚类的方法主要有统计学方法（基于几何距离的聚类）、机器学习（无监督归纳的概念聚类方法）、神经网络方法（无监督学习方式的自组织神经网络）和面向数据库的方法（基于随机搜索的、聚焦的以及聚类特征树等）。

3. 数据归纳 是一种可以抽取数据子集的简洁模式。有两类方法可用于数据归纳，一种是从水平方向（事例）上归纳数据，另一种是从垂直方向（字段）上归纳数据。对于前者可产生子集的摘要，如足够的统计数据基础上的归纳或生成支持子集成立的逻辑条件。对于后者，可在字段中预测关系。一种常见的数据归纳方法称为关联规则。关联规则表明了在确定的频度和支持度上，一定的值的组合发生在另一些值的组合的基础之上。关联规则的一个应用就是超市货篮分析。

4. 依赖模型化 是在给定数据中获取一些偶然的结构。这种偶然的模型基于一定的概率分布或者基于数据集的字段间的确定的函数依赖。一般的密度估计方法属于这种类型。

5. 发现变化和偏差 这种技术面向的是序列信息，如时序或其它次序，这类方法的显著特征是所观察到的现象的次序是重要的。数据库中发现频繁序列的尺度化方法可在实时事物数据库中进行数据挖掘，其复杂度最坏情况下是指数级的。

3.1.4 计算智能方法在数据挖掘中的应用与研究

因为数据挖掘是从大量数据中识别出有效的、新颖的、潜在有用的，以及最终可理解的知识和模式的高级操作过程，所以数据挖掘也可以说是一个模式识别的过程，因此许多模式识别领域的技术经过一定的改进便可以在数据挖掘中起重要的作用。计算智能（Computational Intelligence ，CI）方法是传统人工智能（Artificial Intelligence，AI）的扩展，它是模式识别技术发展的新阶段，因此，CI技术在数据挖掘算法中的应用是未来 KDD 研究的主要方向之一。

1.神经网络方法在数据挖掘中的应用

在 KDD 领域中，应用比较广泛的人工神经网络模型有：BP 网、SOM（Self-Organizing Feature Map）网（包括 Kohonen 和 ART 模型）、循环 BP 网、RBF（Radial Basis Function）网和 PNN（Probabilistic Neural Networks）网等。BP 网采用多层前向的拓扑形状，有输入层、中间层和输出层组成，与传统的统计分析比较，它可以同时逼近多个输出，以满足多输入参数所造成的复杂情况并且实现简单。BP 网可以被用于分类、回归和时间序列预测等 KDD 任务中。另外，如果 KDD 的目标是对时间序列进行预测时，用循环 BP 网比普通的 BP 网要好；需

要在线学习时，ART 和 RBF 的训练数据相对快一些。

KDD 系统的目标是通过分析和预测从数据库中抽取出特定的知识，因此，在大多数情况下，对数据对象进行模式识别的规则是预先未知的，聚类方法也就成为数据挖掘算法的核心。SOM 神经网络模型适合对数据对象进行聚类，SOM 方法是一种基于欧氏距离反复地进行聚类和聚类中心修改的过程，因此，它主要针对与数据库中数据对象的数值属性。

在数据挖掘领域中，神经网络的模型描述能力强；精确性和鲁棒性较好；并且一般不需要专家先验的主观知识支持。但它需要对数据作多遍扫描，训练时间较长，可伸缩性差。由于知识是以网络结构和连接权值的形式来表达的，因此数据挖掘所得到的知识可理解性及开放性都很差。

2.遗传算法在数据挖掘中的应用

遗传算法（Genetic Algorithm ，GA）是一种抽象于生物进化过程的、基于自然选择和生物遗传机制的优化技术，它模仿生物通过染色体的交叉及其基因的遗传变异机制来达到自适应优化的效果，它具有方便实用、便于并行处理、鲁棒性强等特点。在最新的 KDD 技术中，常将其应用于规则的优化计算和分类机器学习。由于各种复杂的结构可以用简单的位串形式进行编码，并且能够不断地用基本变换来改进这些结构，因此 GA 可以在搜索空间（KDD 中的规则空间）中收敛到全局最优解。GA 在数据挖掘中的应用过程如下：

step1：把规则空间通过编码映射到遗传空间中的染色体子空间去。

step2：根据问题和染色体编码构造适应度函数 f（X），一般情况下，适应度函数 f（X）的值越大，X 越接近问题的解。

step3：根据问题和可行性解的编码结构确定遗传算子（Selection, Crossover and Mutation operator）。

step4：种群在遗传算子的作用下产生新的群体。

step5：判断新的群体中是否存在问题的最优解，如存在则输出结果，否则转向 step2。

正是基于 CI 自适应、自组织学习的特点，它已成为很多特殊领域中数据挖掘的有力工具。

3.2 关联规则挖掘

关联规则挖掘用于寻找给定数据集中项之间的有趣的关联或相关关系。关联规则揭示了数据项间的未知的依赖关系，根据所挖掘的关联关系，可以从一个数据对象的信息来推断另一个数据对象的信息。

关联规则的一个典型例子是购物篮分析，系统通过对顾客放入其购物篮中的

不同商品的分析，了解顾客的购买习惯及行为特征。例如，在一次购物消费中，如果顾客购买牛奶的同时，也购买面包的可能性有多大？关联规则的挖掘通过规则的支持度和置信度进行兴趣度度量，这两种度量反映了所发现规则的有用性和确定性。一个关联规则是有趣的，意味着它满足最小支持度阈值和最小置信度阈值。阈值由领域专家和用户设定。一旦发现了有趣的规则，可以帮助零售商有选择地推销，从而引导消费。

事实上，关联规则的挖掘也可用于分析 *Ineternet* 用户的浏览习惯、关联行为等。

3.2.1 关联规则的基本概念

设 $I=\{i_1, i_2, \ldots, i_m\}$是二进制文字的集合，其中的元素称为项(item)。记 D 为交易(transaction)T 的集合，这里交易 T 是项的集合，并且 $T \subseteq I$ 。对应每一个交易有唯一的标识，如交易号，记作 TID。设 X 是一个 I 中项的集合，如果 $X \subseteq T$，那么称交易 T 包含 X。

一个关联规则是形如 $X \Rightarrow Y$ 的蕴涵式，这里 $X \subset I$，$Y \subset I$，并且 $X \cap Y=\Phi$。规则 $X \Rightarrow Y$ 在交易数据库 D 中的支持度（*support*）是交易集中包含 X 和 Y 的交易数与所有交易数之比，记为 support($X \Rightarrow Y$)，即

$$support(X \Rightarrow Y)=|\{T:X \cup Y \subseteq T，T \in D\}|/|D|$$

规则 $X \Rightarrow Y$ 在交易集中的可信度（*confidence*）是指包含 X 和 Y 的交易数与包含 X 的交易数之比，记为 confidence($X \Rightarrow Y$)，即

$$confidence(X \Rightarrow Y)=|\{T: X \cup Y \subseteq T，T \in D\}|/|\{T:X \subseteq T，T \in D\}|$$

给定一个交易集 D，挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度(minsupp)和最小可信度(minconf)的关联规则。

3.2.2 关联规则的种类

我们将关联规则按不同的情况进行分类：

1．基于规则中处理的变量的类别，关联规则可以分为布尔型和数值型。

布尔型关联规则处理的值都是离散的、种类化的，它显示了这些变量之间的关系；而数值型关联规则可以和多维关联或多层关联规则结合起来，对数值型字段进行处理，将其进行动态的分割，或者直接对原始的数据进行处理，当然数值型关联规则中也可以包含种类变量。

例如：性别=“女”=>职业=“秘书”，是布尔型关联规则；性别=“女”=>avg（收入）=2300，涉及的收入是数值类型，所以是一个数值型关联规则。

2.基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

在单层的关联规则中，所有的变量都没有考虑到现实的数据是具有多个不同的层次的；而在多层的关联规则中，对数据的多层性已经进行了充分的考虑。例如：IBM 台式机=>Sony 打印机，是一个细节数据上的单层关联规则；台式机=>Sony 打印机，是一个较高层次和细节层次之间的多层关联规则。

3.基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。

在单维的关联规则中，我们只涉及到数据的一个维，如用户购买的物品；而在多维的关联规则中，要处理的数据将会涉及多个维。换成另一句话，单维关联规则是处理单个属性中的一些关系；多维关联规则是处理各个属性之间的某些关系。

例如：啤酒=>尿布，这条规则只涉及到用户的购买的物品；性别="女"=>职业="秘书"，这条规则就涉及到两个字段的信息，是两个维上的一条关联规则。

给出了关联规则的分类之后，在下面的分析过程中，我们就可以考虑某个具体的方法适用于哪一类规则的挖掘，某类规则又可以用哪些不同的方法进行处理。

3.2.3 关联规则挖掘的算法

3.2.3.1 经典频集方法

Agrawal 等于 1993 年[46]首先提出了挖掘顾客交易数据库中项集间的关联规则问题，其核心方法是基于频集理论的递推方法。以后诸多的研究人员对关联规则的挖掘问题进行了大量的研究。他们的工作包括对原有的算法进行优化，如引入随机采样、并行的思想等，以提高算法挖掘规则的效率；提出各种变体，如泛化的关联规则、周期关联规则等，对关联规则的应用进行推广。

1. 核心算法

关联规则挖掘的经典算法是 Agrawal 等[51]在 1993 年提出的 Apriori 算法，它是一种最有影响的挖掘布尔关联规则频繁项集的算法。算法的名字基于这样的事实：算法使用频繁项集性质的先验知识。Apriori 算法使用一种称作逐层搜索的迭代方法，k-项集勇于探索(k+1)-项集。首先，找出频繁 1-项集的集合。该集合记做 L_1。L_1用于找频繁 2-项集的集合 L_2，如此下去，直到不能找到频繁 k-项集。找每个 L_k 需要一次数据库扫描。

2. 频集算法的几种优化方法

虽然 Apriori 算法自身已经进行了一定的优化，但是在实际的应用中，还是存在不令人满意的地方，于是人们相继提出了一些优化的方法。

1.基于划分的方法。Savasere 等[56]设计了一个基于划分(partition)的算法，这个算法先把数据库从逻辑上分成几个互不相交的块，每次单独考虑一个分块并对它生成所有的频集，然后把产生的频集合并，用来生成所有可能的频集，最后计算

这些项集的支持度。这里分块的大小选择要使得每个分块可以被放入主存，每个阶段只需被扫描一次。而算法的正确性是由每一个可能的频集至少在某一个分块中是频集保证的。上面所讨论的算法是可以高度并行的，可以把每一分块分别分配给某一个处理器生成频集。产生频集的每一个循环结束后，处理器之间进行通信来产生全局的候选 k-项集。通常这里的通信过程是算法执行时间的主要瓶颈；而另一方面，每个独立的处理器生成频集的时间也是一个瓶颈。其他的方法还有在多处理器之间共享一个杂凑树来产生频集。更多的关于生成频集的并行化方法可以在[47,55,61]中找到。

2.基于 hash 的方法。一个高效地产生频集的基于杂凑(hash)的算法由 Park 等[54]提出来。通过实验我们可以发现寻找频集主要的计算是在生成频繁 2-项集 L_k 上，Park 等就是利用了这个性质引入杂凑技术来改进产生频繁 2-项集的方法。

3.基于采样的方法。基于前一遍扫描得到的信息，对此仔细地作组合分析，可以得到一个改进的算法，Mannila 等[53]先考虑了这一点，他们认为采样是发现规则的一个有效途径。随后又由 Toivonen[60]进一步发展了这个思想，先使用从数据库中抽取出来的采样得到一些在整个数据库中可能成立的规则，然后对数据库的剩余部分验证这个结果。Toivonen 的算法相当简单并显著地减少了 I/O 代价，但是一个很大的缺点就是产生的结果不精确，即存在所谓的数据扭曲(data skew)。分布在同一页面上的数据时常是高度相关的，可能不能表示整个数据库中模式的分布，由此而导致的是采样 5%的交易数据所花费的代价可能同扫描一遍数据库相近。Lin 和 Dunham 在[52]中讨论了反扭曲(Anti-skew)算法来挖掘关联规则，在那里他们引入的技术使得扫描数据库的次数少于 2 次，算法使用了一个采样处理来收集有关数据的次数来减少扫描遍数。

Brin 等[49]提出的算法使用比传统算法少的扫描遍数来发现频集，同时比基于采样的方法使用更少的候选集，这些改进了算法在低层的效率。具体的考虑是，在计算 k-项集时，一旦我们认为某个(k+1)-项集可能是频集时，就并行地计算这个(k+1)-项集的支持度，算法需要的总的扫描次数通常少于最大的频集的项数。这里他们也使用了杂凑技术，并提出产生“相关规则”（Correlation Rules）的一个新方法，这是基于他们的[48]工作基础上的。

4.减少交易的个数。减少用于未来扫描的事务集的大小。一个基本的原理就是当一个事务不包含长度为 k 的大项集，则必然不包含长度为 k+1 的大项集。从而我们就可以将这些事务移去，这样在下一遍的扫描中就可以要进行扫描的事务集的个数。这个就是 AprioriTid 的基本思想。

3.2.3.2 其他的频集挖掘方法

上面我们介绍的都是基于 Apriori 的频集方法。即使进行了优化，但是 Apriori 方法一些固有的缺陷还是无法克服：

1.可能产生大量的候选集。当长度为 1 的频集有 10000 个的时候，长度为 2 的候选集个数将会超过 10M。还有就是如果要生成一个很长的规则的时候，要产生的中间元素也是巨大量的。

2.无法对稀有信息进行分析。由于频集使用了参数 *minsup*，所以就无法对小于 *minsup* 的事件进行分析；而如果将 *minsup* 设成一个很低的值，那么算法的效率就成了一个很难处理的问题。

下面将介绍两种方法，分别用于解决以上两个问题。

在[62]中提到了解决问题 1 的一种方法。采用了一种 FP-growth 的方法。他们采用了分而治之的策略：在经过了第一次的扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息。随后我们再将 FP-tree 分化成一些条件库，每个库和一个长度为 1 的频集相关。然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法,使得一个 FP-tree 可以放入主存中。实验表明，FP-growth 对不同长度的规则都有很好的适应性，同时在效率上较之 apriori 算法有巨大的提高。

第二个问题是基于这个的一个想法：apriori 算法得出的关系都是频繁出现的，但是在实际的应用中，我们可能需要寻找一些高度相关的元素，即使这些元素不是频繁出现的。在 apriori 算法中，起决定作用的是支持度，而我们现在将把可信度放在第一位，挖掘一些具有非常高可信度的规则。在[63]中介绍了对于这个问题的一个解决方法。整个算法基本上分成三个步骤：计算特征、生成候选集、过滤候选集。在三个步骤中，关键的地方就是在计算特征时 Hash 方法的使用。在考虑方法的时候，有几个衡量好坏的指数：时空效率、错误率和遗漏率。基本的方法有两类：Min_Hashing(MH)和 Locality_Sensitive_Hashing(LSH)。Min_Hashing 的基本想法是：将一条记录中的头 k 个为 1 的字段的位置作为一个 Hash 函数。Locality_Sentitive_Hashing 的基本想法是：将整个数据库用一种基于概率的方法进行分类，使得相似的列在一起的可能性更大，不相似的列在一起的可能性较小。我们再对这两个方法比较一下。MH 的遗漏率为零，错误率可以由 k 严格控制，但是时空效率相对的较差。LSH 的遗漏率和错误率是无法同时降低的，但是它的时空效率却相对的好很多。所以应该视具体的情况而定。最后的实验数据也说明这种方法的确能产生一些有用的规则。

3.2.4 多层和多维关联规则的挖掘

随着数据仓库和 OLAP 技术研究的深入，可以预见大量的数据将经过整合、预处理，从而存入数据仓库之中。在当前，大多数的数据仓库的应用都是进行统计、建立多维以及 OLAP 的分析工作。随着数据挖掘研究的深入，已经有了 OLAP

和数据挖掘相结合的方法[64,65]。

首先一个有效的数据挖掘方法应该可以进行探索性的数据分析。用户往往希望能在数据库中穿行，选择各种相关的数据，在不同的细节层次上进行分析，以各种不同的形式呈现知识。基于 OLAP 的挖掘就可以提供在不同数据集、不同的细节上的挖掘，可以进行切片、切块、展开、过滤等各种对规则的操作。然后再加上一些可视化的工具，就能大大的提高数据挖掘的灵活性和能力。接着，我们来看一下多层和多维关联规则的定义。

多层关联规则：

对于很多的应用来说，由于数据分布的分散性，所以很难在数据最细节的层次上发现一些强关联规则。当我们引入概念层次后，就可以在较高的层次上进行挖掘。虽然较高层次上得出的规则可能是更普通的信息，但是对于一个用户来说是普通的信息，对于另一个用户却未必如此。所以数据挖掘应该提供这样一种在多个层次上进行挖掘的功能。

多层关联规则的分类：根据规则中涉及到的层次，多层关联规则可以分为同层关联规则和层间关联规则。

多层关联规则的挖掘基本上可以沿用“支持度-可信度”的框架。不过，在支持度设置的问题上有一些要考虑的东西。

同层关联规则可以采用两种支持度策略：

1.统一的最小支持度。对于不同的层次，都使用同一个最小支持度。这样对于用户和算法实现来说都比较的容易，但是弊端也是显然的。

2.递减的最小支持度。每个层次都有不同的最小支持度，较低层次的最小支持度相对较小。同时还可以利用上层挖掘得到的信息进行一些过滤的工作。

层间关联规则考虑最小支持度的时候，应该根据较低层次的最小支持度来定。

多维关联规则：

以上我们研究的基本上都是同一个字段的值之间的关系，比如用户购买的物品。用多维数据库的语言就是单维或者叫维内的关联规则，这些规则一般都是在交易数据库中挖掘的。但是对于多维数据库而言，还有一类多维的关联规则。例如：

年龄（X，“20...30”）∧职业（X,“学生”）==> 购买(X，“笔记本电脑”)

在这里我们就涉及到三个维上的数据：年龄、职业、购买。

根据是否允许同一个维重复出现，可以又细分为维间的关联规则（不允许维重复出现）和混合维关联规则（允许维在规则的左右同时出现）。

年龄（X，“20...30”）∧购买(X，“笔记本电脑”)==> 购买(X，“打印机”)

这个规则就是混合维关联规则。

在挖掘维间关联规则和混合维关联规则的时候，还要考虑不同的字段种类：种

类型和数值型。

对于种类型的字段，原先的算法都可以处理。而对于数值型的字段，需要进行一定的处理之后才可以进行。处理数值型字段的方法基本上有以下几种：

1.数值字段被分成一些预定义的层次结构。这些区间都是由用户预先定义的。得出的规则也叫做静态数量关联规则。

2.数值字段根据数据的分布分成了一些布尔字段。每个布尔字段都表示一个数值字段的区间，落在其中则为 1，反之为 0。这种分法是动态的。得出的规则叫布尔数量关联规则。

3.数值字段被分成一些能体现它含义的区间。它考虑了数据之间的距离的因素。得出的规则叫基于距离的关联规则。

4.直接用数值字段中的原始数据进行分析。使用一些统计的方法对数值字段的值进行分析，并且结合多层关联规则的概念，在多个层次之间进行比较从而得出一些有用的规则。得出的规则叫多层数量关联规则。

在 OLAP 中挖掘多层、多维的关联规则是一个很自然的过程。因为 OLAP 本身的基础就是一个多层多维分析的工具，只是在没有使用数据挖掘技术之前，OLAP 只能做一些简单的统计，而不能发现其中一些深层次的有关系的规则。当我们将 OLAP 和 DataMining 技术结合在一起就形成了一个新的体系 OLAM(On-Line Analytical Mining) [64]。

3.2.5 关联规则价值衡量的方法

当我们用数据挖掘的算法得出了一些结果之后，数据挖掘系统如何知道哪些规则对于用户来说是有用的、有价值的？这里有两个层面：用户主观的层面和系统客观的层面。

3.2.5.1 系统客观层面：

很多的算法都使用“支持度-可信度”的框架。这样的结构有时会产生一些错误的结果。看如下的一个例子：

假设一个提供早餐的零售商调查了 4000 名学生在早晨进行什么运动，得到的结果是 2200 名学生打篮球，2750 名学生晨跑，1800 名学生打篮球、晨跑。那么如果设 minsup 为 40%，minconf 为 60%，我们可以得到如下的关联规则：

打篮球 ⇒ 晨跑 （1）

这条规则其实是错误的，因为晨跑的学生的比例是 68%，甚至大于 60%。然而打篮球和晨跑可能是否定关联的，即当我们考虑如下的关联时：

打篮球 ⇒（不）晨跑 （2）

虽然这条规则的支持度和可信度都比那条描述正向关联的规则（1）低，但是

它更精确。然而，如果我们把支持度和可信度设得足够低，那么我们将得到两条矛盾的规则。但另一方面，如果我们把那些参数设得足够高，我们只能得到不精确的规则。总之，没有一对支持度和可信度的组合可以产生完全正确的关联。

于是人们引入了兴趣度，用来修剪无趣的规则，即避免生成“错觉”的关联规则。一般一条规则的兴趣度是在基于统计独立性假设下真正的强度与期望的强度之比，然而在许多应用中已发现，只要人们仍把支持度作为最初的项集产生的主要决定因素，那么要么把支持度设得足够低以使得不丢失任何有意义的规则，或者冒丢失一些重要规则的风险；对前一种情形计算效率是个问题，而后一种情形则有可能丢失从用户观点来看是有意义的规则的问题。

在[56]中作者给出了感兴趣的规则的定义（R-interesting），在[57]中他们又对此作了改进。在[54]中把事件依赖性的统计定义扩展到兴趣度的定义上来；[59]定义了否定关联规则的兴趣度。

除了把兴趣度作为修剪无价值规则的工具，现在已有许多其他的工作来重新认识项集，如 Brin 等[48]考虑的相关规则。在[49]中讨论了蕴涵规则(implication rule)，规则的蕴涵强度在[0,∞]之间变化，其中蕴涵强度为 1 表示完全无关的规则，∞表示完备的规则，如果蕴涵强度大于 1 则表示更大的期望存在性。

另一个度量值——“收集强度”(collective strength)在[66]中被定义，他们设想使用“大于期望值”来发现有意义的关联规则。项集的“收集强度”是[0,∞]之间的一个数值，其中 0 表示完备的否定相关性，而值∞表示完备的正相关性。详细的讨论可以在[54]中找到。

3.2.5.2 用户主观层面：

上面的讨论只是基于系统方面的考虑，但是一个规则的有用与否最终取决于用户的感觉。只有用户可以决定规则的有效性、可行性。所以我们应该将用户的需求和系统更加紧密的结合起来。

可以采用一种基于约束(consraint-based)[65]的挖掘。具体约束的内容可以有：

1.数据约束。用户可以指定对哪些数据进行挖掘，而不一定是全部的数据。

2.指定挖掘的维和层次。用户可以指定对数据哪些维以及这些维上的哪些层次进行挖掘。

3.规则约束。可以指定哪些类型的规则是我们所需要的。引入一个模板（template）的概念，用户使用它来确定哪些规则是令人感兴趣的而哪些则不然：如果一条规则匹配一个包含的模板（inclusive template），则是令人感兴趣的，然而如果一条规则匹配一个限制的模板（rextrictive template），则被认为是缺乏兴趣的。

其中有些条件可以和算法紧密的结合，从而即提高了效率，又使挖掘的目的更加的明确化了。其他的方法还有：

Kleinberg 等人的工作是希望建立一套理论来判断所得模式的价值，他们认为

这个问题仅能在微观经济学框架里被解决，他们的模型提出了一个可以发展的方向。他们引入并研究了一个新的优化问题——分段(Segmentation)问题，这个框架包含了一些标准的组合分类问题。这个模型根据基本的目标函数，对“被挖掘的数据”的价值提供一个特殊的算法的视角，显示了从这方面导出的具体的优化问题的广泛的应用领域。

在[50]中 Korn 等就利用猜测误差(这里他们使用“均方根”来定义)来作为一些从给定的数据集所发现的规则的“好处”(goodness)的度量，他们所定义的比例规则就是如下的规则：

顾客大多数分别花费 1:2:5 的钱在“面包”:“牛奶”:“奶油”上

通过确定未知的(等价的，被隐藏的，丢失的)值，比例规则可以用来作决策支持。如果数据点线性地相关的话，那么比例规则能达到更紧凑的描述，即关联规则更好地描述了相关性。

3.3 本章小结

本章讨论了数据挖掘技术的背景、任务、分类及计算智能方法在这一领域的应用。作为一门崭新的学科，各种各样的新技术被广泛应用在数据挖掘的各个阶段。人们通过这些分析工具，可以从海量的数据中找到自己需要的相关信息。另外，我们介绍了关联规则挖掘的相关概念，了解了多层关联规则和多维关联规则的基本定义，我们还提出了挖掘关联规则的几种常用算法，以及如何衡量关联规则的价值方法。

第四章　基于多克隆选择的多维关联规则挖掘算法

4.1 关联规则的一些定义[1]

给定一个事务(交易)数据库，人们往往希望发现事务中的关联事实，即事务中一些项目的出现必定隐含着同次事务中其它项目的出现，这是关联规则的一个简单的描述。

设 D 是事务数据库，$I=\{i_1, i_2, \ldots, i_m\}$是所有项目的集合，其中 i_j 是一个项目。每个事务 T_i 是项集，$T_i \subseteq I$，标识符为 TID_i。

定义 4.1.1 设 A，B 是项集，蕴涵式 $A \Rightarrow B$ 称规则，其中 $A \subset I$，$B \subset I$，且 $A \cap B = \Phi$。

定义 4.1.2 设 D 是事务集，A，B 为项集，且有规则 $A \Rightarrow B$。如果 D 中，包含 $A \cup B$ 事务所占比例为 $s\%$，称 $A \Rightarrow B$ 有支持度（*support*）s，即概率 $P(A \cup B)$。

定义 4.1.3 设 D 是事务集，A，B 为项集，且有规则 $A \Rightarrow B$。如果 D 中，$c\%$的事务包含 A 的同时也包含 B，则称 $A \Rightarrow B$ 有置信度（*confidence*）c，即条件概率 $P(B \mid A)$。

这里，不考虑项目在事务中出现的次数。

项集的支持度也是指包含该项目集的事务在 D 中所占的比例。置信度表明了蕴涵的强度，而支持度则表明了 $A \Rightarrow B$ 模式发生的频率。

定义 4.1.4 设 D 是事务集，A，B 为项集，如果 D 中 $e\%$的事务包含事务 B，则称 $A \Rightarrow B$ 有期望置信度（*expected confidence*）e，即概率 $P(B)$。

定义 4.1.5 置信度与期望置信度之比称为作用度（*lift*），其概率表示为 $P(B \mid A) / P(B)$。

显然，置信度是对关联规则的准确度的度量，而支持度则是对关联规则的重要性的度量，期望置信度说明了在有事务集 A 的作用下，对事务集 B 本身的支持度，而作用度则说明了事务集 A 对事务集 B 的影响力的大小。一般地，有用的关联规则的作用度都大于 1。置信度、支持度、期望置信度和作用度等四个参数中，常用的是置信度和支持度。下面以这两个参数为讨论的依据。

定义 4.1.6 设 D 是事务集，A，B 为项集，若 $A \Rightarrow B$ 满足置信度 c 和支持度 s，则称 $A \Rightarrow B$ 为关联规则。

对关联规则 $A \Rightarrow B$，若同时满足最小支持度阈值和最小置信度阈值，则称其为强规则。

一般地，由用户给定最小置信度和最小支持度，发现关联规则的任务就是从数据库中发现那些置信度和支持度都大于给定阈值的强规则，也就是说，挖掘相关规则的关键是在大型数据库中发现强规则。

定义 4.1.7 项的集合称为项集。包含 K 个项的项集称为 K—项集。项集的出现频率是包含项集的事务数，称项集的频率、支持计数或计数。项集满足最小支持度是指如果项集的出现频率大于或等于最小支持度与 D 中事务总数的乘积。如果项集满足最小支持度，则称它为频繁项集。

4.2 Apriori 算法

挖掘关联规则的经典算法是 Apriori 算法[1]，这是一个基于两阶段频集思想的方法，将关联规则挖掘算法的设计可以分解为两个子问题：

1.找到所有支持度大于最小支持度的项集（Itemset），这些项集称为频集（Frequent Itemset)。

2.使用第 1 步找到的频集产生期望的规则。

这里的第 2 步相对简单一点。如给定了一个频集 $Y=I_1I_2...I_k$，$k\geq 2$，$I_j\in I$，产生只包含集合$\{I_1, I_2, ..., I_k\}$中的项的所有规则(最多 k 条)，其中每一条规则的右部只有一项，(即形如$[Y-I_i]\Rightarrow I_i$，$\forall 1\leq i\leq k$)，这里采用的是[4]中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。对于规则右部含两个以上项的规则，在其以后的工作中进行了研究，本文后面考虑的是这种情况。

为了生成所有频集，使用了递推的方法。其核心思想如下：

```
(1)   L1 = {large 1-itemsets};
(2)   for (k=2; Lk-1≠Φ; k++) do begin
(3)       Ck=apriori-gen(Lk-1);    //新的候选集
(4)       for all transactions t∈D do begin
(5)           Ct=subset(Ck,t);     //事务 t 中包含的候选集
(6)        for all candidates c∈ Ct  do
(7)        c.count++;
(8)       end
(9)      Lk={c∈ Ck |c.count≥minsup}
(10)  end
(11)  Answer=∪kLk;
```

首先产生频繁 1-项集 L_1，然后是频繁 2-项集 L_2，直到有某个 r 值使得 L_r 为空，这时算法停止。这里在第 k 次循环中，过程先产生候选 k-项集的集合 C_k，C_k 中的每一个项集是对两个只有一个项不同的属于 L_{k-1} 的频集做一个(k-2)-连接来产生的。C_k 中的项集是用来产生频集的候选集，最后的频集 L_k 必须是 C_k 的一个子集。C_k 中的每个元素需在交易数据库中进行验证来决定其是否加入 L_k，这里的验证过

程是算法性能的一个瓶颈。这个方法要求多次扫描可能很大的交易数据库，即如果频集最多包含 10 个项，那么就需要扫描交易数据库 10 遍，这需要很大的 I/O 负载。

Agrawal 等引入了修剪技术（Pruning）来减小候选集 C_k 的大小，由此可以显著地改进生成所有频集算法的性能。算法中引入的修剪策略基于这样一个性质：一个项集是频集当且仅当它的所有子集都是频集。那么，如果 C_k 中某个候选项集有一个(k-1)-子集不属于 L_{k-1}，则这个项集可以被修剪掉不再被考虑，这个修剪过程可以降低计算所有的候选集的支持度的代价。文[51]中，还引入杂凑树（Hash Tree）方法来有效地计算每个项集的支持度。

传统的 Apriori 算法要产生大量的候选项集，且每次都需要重复的扫描数据库，通过模式匹配检查一个很大的候选集合。改进后的 Apriori 算法虽然克服了传统算法的一些缺点，但仍然存在挖掘速度慢、计算过程复杂、不能并行化处理等问题。所以就要求我们设计出一种的新的方法来进行关联规则挖掘，特别要针对多维关联规则挖掘问题提出新的具有较好性能的算法。

4.3 基于多克隆选择算法的多维关联规则挖掘算法

通过对关联规则挖掘的深入研究，我们发现在应用中急需一种能兼顾适应度和支持度条件、同时挖掘出多个关联规则的快速算法。而多克隆选择算法恰恰符合这一条件，它的收敛速度快，具有并行性和记忆功能，并且不会导致种群多样性的减弱，具有很强的全局及局部搜索能力，故而我们提出了基于多克隆选择的多维关联规则挖掘算法。

4.3.1 染色体的编码

多克隆选择算法建立在编码的基础之上，合适的编码方法会提高后续工作的效率。在此，我们采用十进制编码。

在关联规则的挖掘中，通过对数据进行概化和归纳，可能会删除一些对数据挖掘没有太大意义的属性列，然而事实表通常仍保留了多个属性列。例如，在一个公司的销售事实表中，可能有客户年龄段、客户收入层次、客户职业、所购买物品等许多属性。但最后挖掘出的关联规则并不一定包含所有的属性值。我们可能会挖掘出形如

age (30+) ^ occuption(worker)=>item_bought (Changhong-TV)

这样的规则，此规则并不包含 income 项。

多维关联规则挖掘所得到的一般是由各个属性的合取式组成的形如 A1^A2^…^An=>B1^B2^…^Bm 的规则，我们可以用这样一个大代码段表示：

a、每个属性对应一个较小的编码段；

b、这些较小的代码段以同一顺序排列成大的代码段；

在实际操作过程中，我们采用实值编码，假设有一个由 age、income、occuption、item_bought 组成的事实表（其中 age 属性有 6 个值、income 有 10 个值、occuption 有 30 个值，item_bought 有 25 个值），我们编码的范围为 0 0 0 0 到 6 10 30 25，其中 0 表示这个属性未被选中。

4.3.2 亲和度函数的构造

亲和度函数 f 是评价抗体与抗原联系的量化反映，它的选取对于克隆算法具有举足轻重的作用。

在关联规则挖掘中，支持度是对关联规则重要性的衡量，它说明了关联规则在所有事物中的代表性，它的大小反映了关联规则在实际应用中普遍性的大小。置信度反映了由相关条件推出结论的正确率，如果置信度达不到一定的阈值，那么这条关联规则就没有意义。所以，我们先选用支持度作为筛选条件，以置信度作为亲和度函数，它可以表示为：

$$f = C$$

其中 C 为置信度。

4.3.3 基于多克隆选择的多维关联规则挖掘算法

步骤一：随机产生每一属性值，以概率 α_i 取 0 选取此属性值，以概率（1-α_i）选取其他属性值，其范围为从 1 到此属性值个数间随机选择的一个整数。当某一属性对应的选取概率 α_i=0 时，此属性一定存在于所挖掘出的关联规则之中，若 α_i 不为 0，则其对应的属性不一定存在于所挖掘出的关联规则之中。所以，我们如果要挖掘出包含特定属性的关联规则时，则应将此属性的选取概率 α_i 取 0，其余属性的选取概率 α_i 一般取 0.2~0.5。循环选取 n 个初始抗体，这些抗体中各个属性的顺序相同，且应保证每个抗体满足支持度阈值条件。由此形成最初的抗体种群 $\overline{A}(k)$。

步骤二：计算出每一抗体的 q_i，对抗体种群进行克隆操作 T_c^C。克隆过后，种群变为 $\overline{A}'(k) = \{\overline{A_1}'(k), \overline{A_2}'(k), \ldots, \overline{A_n}'(k)\}$。

步骤三： 对目前种群 $\overline{A}'(k)$ 进行克隆变异操作 $\overline{A}''(k) = T_m^c(\overline{A}'(k))$，我们以概率 P_m^c 从 $\overline{A}'(k)$ 中抽出抗体，对一个或多个属性进行实值变异，使其以一定概率随机变为其他属性值，删去此种群中不满足支持度条件的抗体。

步骤四：对目前种群 $\overline{A}''(k)$ 进行克隆交叉操作 $\overline{A}'''(k) = T_x^c(\overline{A}''(k))$，交叉时我们使用离散重组法则，删去此种群中不满足支持度条件的抗体。

步骤五：对目前种群 $\overline{A}'''(k)$ 进行克隆选择操作 $\overline{A}(k+1) = T_s^c(\overline{A}'''(k))$。若得到的某

个抗体同时满足最小支持度和最小置信度条件，则输出此抗体，并把此抗体还原为原始属性值，但仍把此抗体保留在种群之中。如果迭代次数满足停机条件，则停机；否则，把此时种群作为下一代计算的初始抗体种群，转步骤二。

4.3.4 仿真试验与结果分析

表 1 是一个经过属性概化后的销售事实表。表 1 数据来源：http://kddforum.126.com

Age	Gender	income	occuption	category	Brand
20...25	Male	20k...29k	student	computer	IBM
25...30	Female	40k...49k	worker	VCR	Sony
20...25	Male	30k...39k	Doctor	TV	Toshiba
15...20	Male	10k...19k	student	walkman	Sony
30...35	Female	40k...49k	manager	computer	Apple
...	...	...	...	...	....

表 4.1 一个经过属性概化后的销售表

如果我们需要挖掘形如 $A_1 \cap A_2 \cap ... \Rightarrow category$ 的关联规则，支持度阈值为 3%、置信度阈值为 60%。则可以取变异概率 P_m 为 0.2,交叉概率 P_c 为 0.5，初始种群 n 选为 300 个，克隆规模 N_c 选为 900 个，迭代次数为 1000 代。编码时 category 列的 α_i 值取 0。则我们可以得到以下规则：

$Age(15...20) \Rightarrow category(walkman)$ [support=6% confidence=75%]

$Age(25...30) \cap Income(40k...49k) \Rightarrow category(VCR)$

[support=4% confidence=65%].....

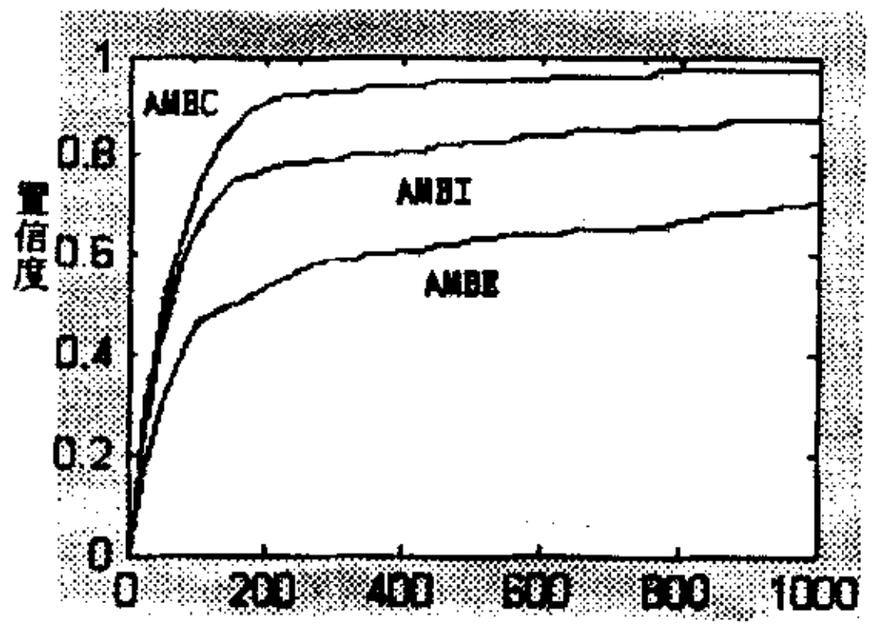
所用算法	挖掘出的规则数	规则的提取率	计算时间
Apriori 算法	12	100%	90.23min
基于进化算法的关联规则挖掘算法[1]	8.54	71.3%	9.01min
基于免疫算法[29]的关联规则挖掘算法	10.43	86.9%	9.23min
基于多克隆算法的多维关联规则挖掘算法	11.63	96.9%	9.16min

表 4.2 几种算法挖掘出的关联规则数、规则提取率、计算时间比较

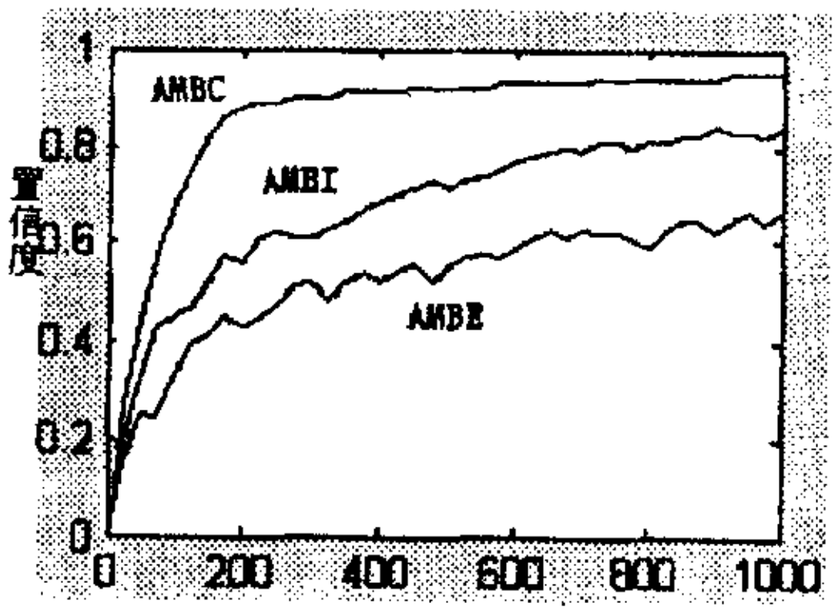
(以上所有取值均为计算 20 次的平均值)

表 2 中规则的提取率为相应算法挖掘出的规则数与总规则数的比率。由表 2 可以看出，传统的 Apriori 算法挖掘出的规则数最多，后三种算法挖掘出的规则较少。但 Apriori 算法在挖掘多维关联规则时运算量太大，且要经过后期置信度的处理，而且不能实现并行化，而后三种方法克服了这些缺点。由图 1 可以看出，基

于多克隆选择的多维关联规则挖掘算法比其他两种算法的执行性能要好，无论是平均置信度还是最佳置信度，都有明显的优势。



AMBC 基于多克隆的多维关联规则挖掘算法
AMBI 基于免疫的关联规则挖掘算法
AMBE 基于进化的关联规则挖掘算法



AMBC 基于多克隆的多维关联规则挖掘算法
AMBI 基于免疫的关联规则挖掘算法
AMBE 基于进化的关联规则挖掘算法

图 4.1 几种关联规则挖掘算法的比较

(a)最佳置信度比较；(b)平均置信度比较

以表 1 为例，如果我们还需要挖掘形如 $A_1 \cap A_2 \cap \ldots \Rightarrow brand$ 和形如 $A_1 \cap A_2 \cap \ldots \Rightarrow category \cap brand$ 的关联规则，那么我们只需使 category 和 brand 列的 α_i 值取适当的数，支持度和置信度作相应的计算，则可以挖掘出上面所有三种格式的关联规则。如果我们需要挖掘出同一属性中的关联规则，则编码时可以重复选取同一个属性构成初始串。如果要选择混合上面两种要求的关联规则，则只需在构造初始串时进行相应的操作即可。

另外，与 Kim, J., Ong, A. 等人提出的基于 CIFD(Computer Immune system for Fraud Detection)的关联规则挖掘算法[67] 相比，由于省去了建立整个 CIFD 的时间，具有更小的计算量，且大大提高了关联规则挖掘的可操作性。

4.4 结论

通过实验我们可以发现基于多克隆选择算法的多维关联规则挖掘算法具有收敛速度快的特点，这样我们就可以适当减少算法的迭代次数，从而提高挖掘效率。其次，此算法同时具有相当好的全局及局部搜索能力，这样可以得到更多的符合条件的关联规则。最后，此算法比起传统的 Apriori 算法具有更好的并行性，这样可以大大加强其搜索能力。

第五章 供应链管理及其相关算法

5.1 供应链的相关概念和分类

5.1.1 供应链定义[68]

所谓供应链(Supply Chain)，是指产品生产和流通过程所涉及的原材料供应商、生产商、批发商、零售商以及最终消费者组成的供需网络，即由物料获取、物料加工、并将成品送到用户手中这一过程所涉及的企业和企业部门组成的一个网络。供应链一般分为内部供应链和外部供应链。

5.1.2 供应链的分类

供应链分为内部供应链和外部供应链。内部供应链是指企业内部产品生产和流通过程中所涉及的采购部门、生产部门、仓储部门、销售部门等组成的供需网络。而外部供应链则是指企业外部的、与企业相关的产品生产和流通过租中涉及的原材料供应商、生产厂商、储运商、零售商以及最终消费者组成的供需网络。内部供应链和外部供应链共同组成了企业产品从原材料到成品到消费者的供应链。可以说，内部供应链是企业外部供应链的缩小化。如对于产品制造业的厂商，其采购部门就可以看做外部供应链中的供应商。它的区别只在于外部供应链范围大，涉及企业多，企业间协调更加困难。

5.1.3 供应链管理的背景

供应链管理，是指人们在认识和掌握了供应链各环节内在规律和相互联系的基础上，利用管理的计划、组织、指挥、协调、控制和激励职能，对产品生产和流通过程中各个环节所涉及的物流、信息流、资金流、业务流等进行合理的调控，以期达到最佳的组合，发挥最大的效率，迅速以最小的成本为客户提供最大的附加值。供应链管理是在现代科技条件下发展起来的管理理念，它涉及企业及企业管理的方方面面，是一种跨行业的管理。

在企业内部供应链管理中，各部门不是再像原来那样只强化自己本部门的目标，而不考虑其他部门的目标或整体目标。他们明白了各个部门的成功不一定导致整个企业的成功，因而转向以追求各部门的总体目标即企业目标为目的。ERP就是管理企业内部供应链以实现这一目标的较好方法之一。

在外部供应链管理中，企业之间作为贸易合作伙伴，也为追求共同经济利益的最大化而共同努力。因为他们清楚，只有提高整个供应链的竞争能力，才能稳固企业自身的竞争能力与市场地位。

以整个供应链的利益最大化为目标是供应链管理的本质。要实现这一目标，

需要电子商务的环境与技术；而搞好供应铭管理，则更深入地推动了电子商务。所以要想真正从本质上理解电子商务并深入开展电子商务，必须加强对供应链的管理。

5.2 供应链管理中的各项技术

在全球供应链中，管理者必须随时准备应付以下情况发生：产品在市场上领先性的丧失，昂贵的运输费用，过高的库存水平，销售预测不准确，以及在解决技术难题中的滞后。所有这些问题的解决都必须要有一套对于整个供应链各个环节的有效而协调的管理策略。

5.2.1 生产/制造系统

这是供应链中必不可少的一环。由原材料的加工，零件的组装与拆装，到最后的成型，包装，都属于其范围。因此，制造系统的优化是供应链的一个关键问题：怎样使一个生产线更具效益，怎样充分利用机器和机器之间的缓冲区，在机器以一定的概率分布损坏或发生故障，以及缓冲区以一定的概率分布阻塞的情况下让生产率达到最优。目前，针对制造系统的优化，给出了几种不同类型的模型：基于数学规划的，基于计算机仿真的，基于 Markov 链的模型等多种形式；但由于排序(Scheduling)问题即使在机器较少的情况下依然是一个 NP-完全问题，而在一个机器和缓冲区众多的车间，要实现生产线的同步高速运行，安排缓冲区的大小以及怎样配置机器的修理工是一个很难的问题，由于几乎无法用解析的方法描述这一问题，因此发展了很多可操作且有效的求近似最优解的算法，目前的寻优算法包括分支定界、遗传算法及 Tabu 搜索等启发式方法。

5.2.2 运输系统和时间表问题

在全球性经济方式的今天，物流信息流等在多地乃至全球范围内流动；运输的费用，如何将各种运输方式合理安排才能做到满足生产加工和顾客服务的需要？其间的协作问题，多结点，多连接及各地不同的经济环境使得运输管理的复杂性增大；各个地理政治地区的不同市场结构也为后勤协作提出了问题。时间表的科学安排可以有效利用时间，降低因信息过期带来的损失。

5.2.3 库存决策

实际上，这是 60 年代以来的研究热点。最初关于供应链的研究多集中在这一主题上。近来日本兴起的“即时传输，低库存和零缺陷”的管理方法产生了极大影响，此后似乎库存问题的研究要淡化了，生产和销售之间的渠道只要做到信息

畅通，传输有效率，库存保持一个极低的水准就可以了。但是，实际问题并不这么简单，库存问题仍然必须加以重点研究。

库存问题研究在需求平稳的前提下的单产品单库存，由多个库存点及多种产品就组成了多阶段库存问题。Rogers 和 Tsubakitani[70]在有投资约束的情况下，对于多库存地点的产品部分的库存水平给出了一个必要条件。计算了最优状态的临界比率 F(Z*)=Pu/(Pu+Po)，其中 F(.)是需求的集中分布函数，Z*是最优库存数量，Pu 和 Po 是存货不足和过量存货的费用。

5.3 一些应用于协作管理的算法

以往供应链管理及其研究一般集中链中各环节内部的有效性管理。随着市场全球化进程的快速推进和竞争压力的增加，供应链各方开始寻找一种供应链可以采用的协作管理方法来进行双优或多方最优博弈，以追求更多的利润。

对于双方博弈的供应链，一般有：买方-卖方协作，生产-分配协作，库存-分配协作。

5.3.1 买方-卖方协作

供应链的开始一端是原材料供应商和子装配系统间的协作，这一部分的成本往往要占到销售值的 50%以上。许多传统的库存模型用于确定买方的最优订货数量，但这种模型忽略了两个机会：一个是在不改变订货策略的前提下削减成本，这可以通过在物流设备和数据交换技术方面的投资进行，如电子数据交换 EDI，这属于策略决策；第二，厂商可以找到对于买方和卖方均为最优的订货数量，双方必须交涉谈判如何分割收益。

Monahan 给出了最优订货数量增加的决定因子：k*=(S2/S1+1)1/2，其中 S2，S1 分别是买主和卖主各自的订货成本[69]。Lee 和 Rosenblatt[71]推广了这个模型，加入了最小边际利润并允许卖主以任意数量订货，给出的模型可以同时找到最优订货数量增加因子 K 和卖主的最优订货数量，它是买主订货数量的一个整数倍 k。他们对单买方和单产品的情形设计了一个算法，用决定收益最大数量折扣价格表来计算最大利润，这种方法基于买方采用 EOQ 策略，并假设订货和持有花费已知，适当的订货数量可以由数量折扣表来得到。Banerjee[72]对于单买主单卖主(生产率有限)发展了一个合作经济大范围模型，在先订货后生产的假设下，给出了求解最优合作生产或订货数量的模型。Goyal[73]放松了最大化的生产假设，认为经济的生产数量应该是买主购买数量的整数倍；Anupindi 和 Akella[74]给出了单买主多卖主情形下三类不同模型的最优订货策略。Kohli 和 Park[75]在假设所有的产品分别进行各自的合作订货的前提下探讨了一种合作订货策略方法来减少一个卖主和一群买

主间的交易费用；Lau 和 Lau[76]对于一个买主和两个卖主的情况应用(Q,R)连续评价系统给出了最优订货策略。

现在从另外一个角度，在通过对偶线性规划的思想说明制造商和供应商之间的关系的基础上，提出买卖博弈协作中的双优策略。制造商和供应商间的关系可以通过线性规划理论中的对偶规划来刻划[77]。

设原规划为

$$Maxc^T x, s1.Ax \le b, x \ge 0$$

对偶规划为

$$Minb^T w, s.t.A^T w \ge c, w \ge 0$$

对应于原规划，考虑一个制造商，他用 m 种资源制造 n 种产品，制造一个单位的产品 j(j=1,...,n)用掉 a_{ij} 个单位的资源 i(i=1,...,m)，他已得到了 i 种资源 b_i 个单位，当前市场上 j 产品的单位价格是 c_j。因此，原问题就是引导制造商制订最优生产计划，以便用有限资源去获得最大销售额。

再考虑对偶规划的情况，假设制造商是从供应厂商得到资源的，制造商想和供应厂商协商资源 i 的单位购买价格 w_i。因此，制造商的目标是以最小的总购买价格 $b^T w$ 去获得资源 b_i(i=1,...,m)，由于市场价格 c_j 和“产品-资源”转化率是关于市场的开放信息，制造商从理论知道，一个理性而精明的供应商应当愿意按下式尽可能多地去索求客户(制造商)：

$$a_{1j}w_1 + a_{2j}w_2 + ... + a_{mj}w_m \ge c_j$$

这样，对偶线性规划使制造商终于得到一个最小费用计划，他的购买价格是精明的供应商所能接受的。以上例子中可以看到，对应于买卖双方，存在一种都可以接受的最佳方案，即达到双优。

5.3.2 生产-分配协作

这种协作的形式多种多样，产品可以是去进行制造，或分送给分配中心，零售商及工厂。同时描述任意两种厂商合作的模型很多，但是,要建立同时描述以上这些协作的模型几乎没有，这是因为，以上几种协作问题都是由各自的库存缓冲区而两两隔离的，无法将其归为一类模型；而不同的部门又会对这些协作关系的任一组活动做出反映。又如运输规划和机器排序，二者都是 NPC 问题，众所周知，NPC 问题是国际性的组合论难题。现在，刻划两两协作关系的模型仍然集中于动态规划，非线性规划，混合整数规划，马尔柯夫链等范畴;多产品的仿真也是目前的解决方法之一。

Chien[78]对于单产品的情形试图发现利润最大的生产和运输数量。每周的需求假设为独立和平稳的，并服从一个已知的概率分布。

Chandra 和 Fisher[79]给出了一个单工厂多顾客多阶段的模型，寻求将生产计划

问题和运输路线问题进行综合研究，并提出了非协作关系下和协作关系下的不同求解方法。在非协作关系中，生产计划问题可以直接求得最优解，而运输路线问题的解则可由启发式算法求出。考虑到库存、库存平衡和运输能力的限制，在每个时期通过把为某一个客户的运输与早些时候为它提供的运输相结合来优化这个关系；而在协作关系中的解则先由非协作关系的解导出，然后联系在生产计划中可能产生的变化进一步求得。通过假设的数据来验证这两个关系，其结果表明：生产和分配协作所占比重增加会使得:生产能力所受到的限制减小、时间范围变宽、库存和安装费用降低。

5.3.3 库存-分配协作

事实上，最初对于供应链管理中的研究就是针对这一关系的，近来的研究多集中于多级库存系统；最近，随着顾客对服务需求的增长，供应链中这种协作的研究更重要了，而且开始了对带有随机因素的多级库存系统的研究。

Clark 和 Scarf[80]1960 年撰写的第一篇论文中给出了一个反复的分解方法以决定系列多级结构的最优策略。之后，Silver 和 Peterson[81]给出了一个公式并讨论了简单的两级库存系统；Muckstadt 和 Roundy[82]对于多级生产或分配网络的近似最优解的搜寻给出了一个有效算法。Lee 和 Billington[83]讨论了供应链库存的低效管理的坏处，并分析了 HP 台式打印机公司的分散化全球供应链的物流情况，他们还特别讨论了基于本地信息的供应链的决策。

5.4 现代供应链的流程构造[68]

5.4.1 供应链管理基本方针的确定

所谓基本方针是指在满足最终顾客需求的基础上如何构筑供应链，以及与什么样的企业形成合作关系。在基本方针确立阶段，最为主要的是整个供应链由谁来领导，与此同时，各参与方如厂商、批发商、零售商各自承担什么样的职责。一般来讲，供应链主要有四种形式：A、厂商、零售企业合作经营型；B、大型零售业主导型；C、信息武装批发业主导型；D、厂商、批发业合作经营型。

在明确职责时，一个重要的问题是主导者的主导权究竟是什么，它能否成为统一整个供应链的关键要素，否则主导权的模糊不清不仅无助于计划、设计的制定与实施，同时也无法维系整个供应链的运转，建立起强有力的管理组织，所以说，主导者和主导权的确立是供应链管理的前提条件。

5.4.2 绘制“供应链物流图”

在彻底分析现状的基础上，找出需要改进的问题。在供应链管理中，对物流

现状的分析不是针对一个企业而言，而是产业全体供应链间整个商品流动过程的情况和时间，即商业在各部门、各企业的滞留时间及在库情况。

这种彻底查明从最低供应商到最终顾客为止整个供应链的存在的问题，并实现流通整体改善的方法主要是供应链物流图。在供应链物流图中考察的问题大致有：调达的滞留时间有多长、中心仓库、站头在库是否过多、从顾客服务的角度考虑批发在库是否过多，此外，流通全过程的物流时间是否过长，以及在库配置方面，在哪个企业的哪个仓库如何配置库存才能更好地为最终用户提供高质量的服务等问题。

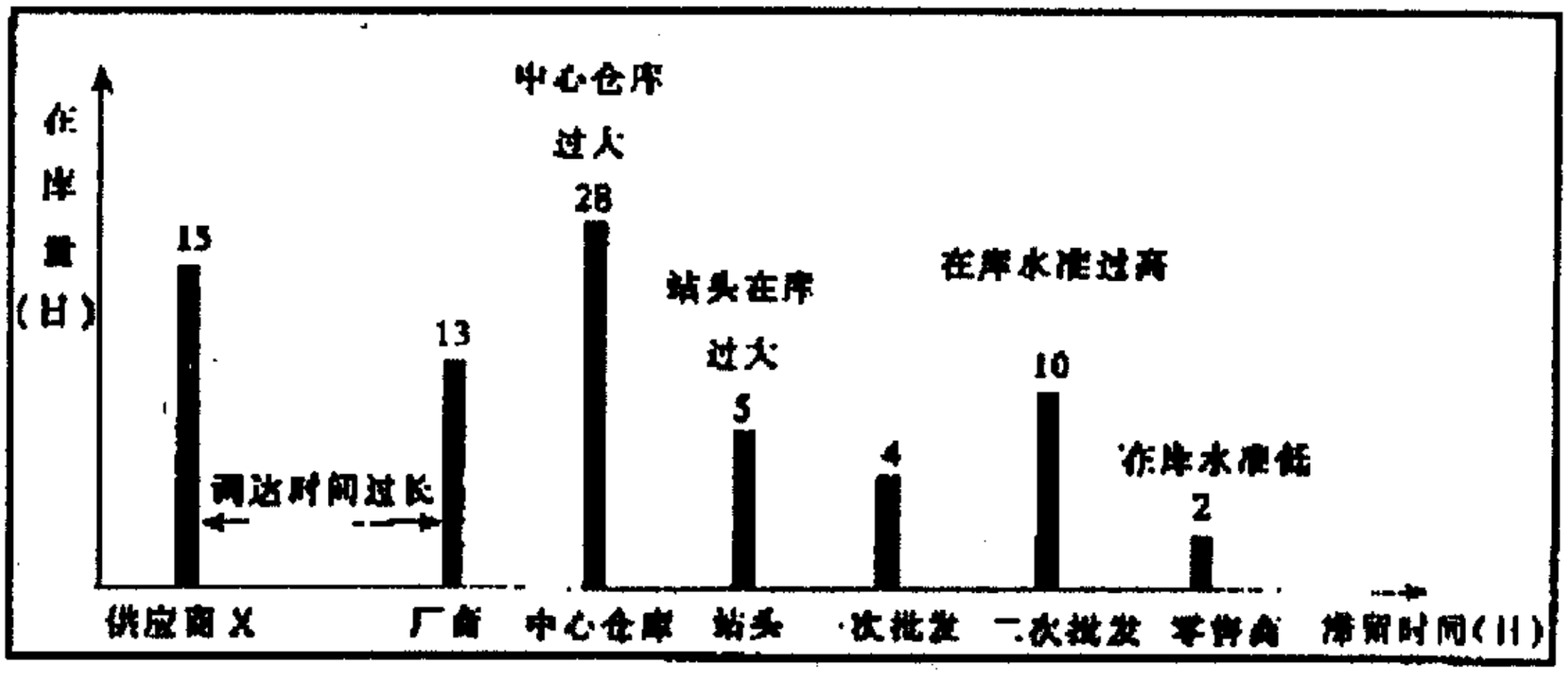


图 5.1 一个供应链物流图

通过供应链物流图能很清楚地知晓商品在整个流通过程中的问题与不足。然而，需要进一步指出的是，分析供应链图，考察物流状况，发现、找出问题不仅仅局限于纵向思维，也需要从横断面来考察供应链，这就要运用定点超越分析法。定点超越(Benchmarking)是市场营销战略规划的一种重要的手段与方法，它是指通过对比本企业与同产业中先进企业或其他产业先进企业在战略设定、计划、实施和控制上的差异，找出本企业的不足与具体差距，并以此作为企业进一步发展超越的目标的一种战略分析方法。这种方法也同样适用于供应链分析，即通过对比先进企业和首位企业的供应链物流图，找出本企业在供应链不同阶段物流时间、在库上的差距，将其作为物流系统改善的目标和方向。

5.4.3 制定具体的供应接管理实施计划

具体制定供应链管理计划主要涉及到管理范围的具体化、管理战略、合作领域、变革领域等等项日。具体看，各项目涵盖的要点有：

1、管理范围

供应链管理范围就是要明确供应链管理纵向和横向的领域。纵向领域是指从供应商到零售商合作对象的选择，即对于某个企业供应链来讲，是否需要厂商、

批发商或零售商的加盟，或者说，本供应链的薄弱环节在什么位置，需要强化的经济主体是哪种类型等等。相对而言，纵向领域较易确定，比较复杂、具有难度的是横向领域的确定。横向领域涉及到流通各环节交易主体的类型、幅度和数量，如零售商是否需要从系列化向非系列化发展，或者从专业店向综合店发展，供应链中的厂商应该是固定的还是竞争开放式的等等问题。显然，这些要素对于供应链长期绩效能产生深远影响，因而是需要慎重考虑的问题。

2、管理战略

管理战略主要涉及到本企业供应链与其他供应链的竞争方法和预期目标，在类型上主要有：

a. 价格破坏型——以追求低价格为主要目标的供应链

b. 竞争生存型——通过各种方式竞争以击败其他对于为目标的供应链

c. 共存型——以追求高附加价值、差异化的流通体系为目标，实现与其他对手共存型的供应链

3、合作领域

可能的合作领域包括物流活动、运输、库存、物流管理(组织、控制等)、流通加工、低成本生产、商品企划、物流网与信息收集等等。

5.4.4 消除供应链间的瓶颈，明确企业活动的规则

在这方面，有两点必须加以关注：

第一，供应链中各合作成员要通过信息公开、共有、计划共有、业务共同化等制度建设，积极地为合作方提供利益。这是供应链制度建设中至关重要的一点。销售、出货、在库信息的非公开往往是供应链的瓶颈，也是合作经营难以开展的主要因素。

第二，合作成员间的风险分担对消除供应链瓶颈具有积极意义。随着 Just-in-time 生产体系的普及以及多额度小单位配送的发展，如何消除由此给零部件供应商和批发商带来的成本上升、风险增大的问题，需要引起重视。只有正确解决了风险分担问题，才能使供应链物流提高效率，取得超越企业、部门界限的合作性利益，相反，一味把风险转嫁出去，只会增加供应链的不稳定。

5.4.5 落实安排业务流程信息系统等企业间具体活动的路径和机构

构筑一定的制度框架和装置是保障供应链管理顺利进行的基础与前提条件。众所周知，从制度经济学的角度看，经济绩效是在一定的制度框架内产生，没有完善的制度装置，将无法保证经济绩效的稳定实现。同样，要发挥供应链在渠道全过程中的优势，实现流通体系的最优化，必须确定供应链建设中相应的路径和装置。

从具体实施情况来看，建立用计算机连接的能够反映物流、信息流的综合系统是供应链管理必不可少的条件(见图 5.2)，亦即在 Pos 信息系统基础上确立各种计划和进货流程。也正因为如此，VAN 和 EDI 的导人、从调达到最终顾客全过程的货物追踪系统和各成员间的构通系统的建立成为供应链构造的重要内容。

从管理机构的设立来讲，各自分散的物流、信息流管理显然不适应供应链管理全过程、整体性的要求，要实现信息的高度集约化，提高决策的判断力，并灵活对应市场变化，必须改变原来垂直型的组织机构，建立兼自律与活用性为一体的网络合作型中间组织，因此，在供应链建设中，往往可以看到厂商、批发商零售商通过合作或合资的形式来建立共同的物流管理机构。

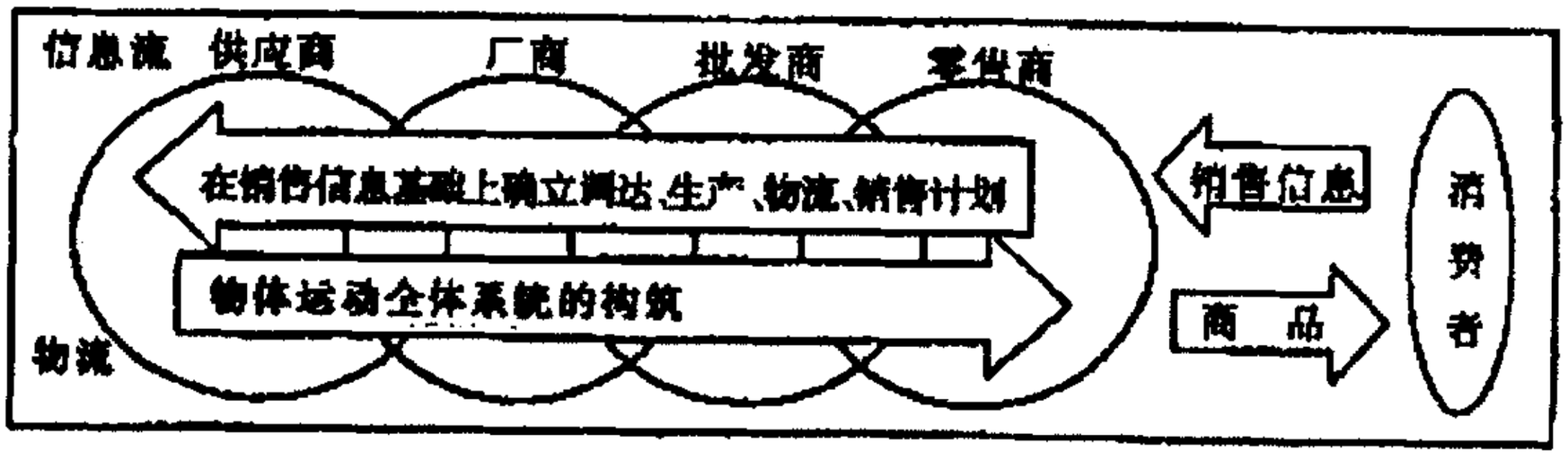


图 5.2 供应链的系统设计

5.5 本章小结

在这一章里，我们简要介绍了供应链管理及其流程构造。

供应链管理是一个新兴学科，它的成功实施将有助于企业获得更多的经济收益，它分为内部供应链和外部供应链。随着供应链管理越来越受到人们的重视，各种各样的方法被应用于供应链管理的各个环节之中。协作管理是供应链管理的一个重要发展方向，人们对协作管理中相关的技术进行了大量的研究，这些研究有助于提高供应链管理的执行效率。

企业如果要在日常操作中享受供应链管理带来的巨大利益，就要构造适合其所属行业的供应链系统，因此采用科学规范的供应链流程构造理论可以帮助企业快速建立其自己的供应链管理系统。

第六章 基于多克隆选择计算的供应链优化算法

6.1 用遗传算法求解供应链模型

供应链模型一般是一个混合整数规划问题，同时它属于一种NP 困难问题，长期以来，人们对它做了很多的研究工作，提出了不少求解该类问题的有价值的方法，如分支定界法、隐枚举法、动态规划法等。它们都在一定程度上求解了该类问题，但也都具有很大的局限性，那就是对于规模较大的问题不易求解，难以在有限时间内找到最优解。

近年来，有研究者开始采用一些近优算法来求解这类混合整数规划问题，如遗传算法。自遗传算法成功地求解了旅行商问题后，它在求解组合优化领域的NP 问题（整数规划就属于这一类NP 问题）上已经显示出了强大的搜索优势。大量实验表明，遗传算法不仅可以获得一个较好的近优解（有时达到最优解），目前已成为人们广泛使用的一种优化方法。

克隆选择算法是一种新的人工免疫算法，它具有进化算法的所有优点，如算法简单，具有固有的并行性等。同时，相比于进化算法，克隆算法具有更多的优点：第一，在算法的实现上，进化算法更多地强调全局搜索，而忽视局部搜索，但克隆选择算法两者兼顾，有更好的种群多样性；第二，克隆选择算法弥补了进化算法较少关注种群间的协作这一缺点，提出了抗体-抗体亲和度概念；第三，克隆算法中变异是主要算子，交叉是次要算子，与进化算法相反，实践中也证明了克隆选择算法的性能强于相应的遗传算法；最后，克隆算子具有记忆性能，因此本身就能保证算法以概率 1 收敛到最优解，而遗传算法则没有这一特性。因此，本文将克隆算法应用于供应链的求解。

6.2 本文采用的供应链模型

本文采用的供应链模型如下[26]：

$$\min Z = \sum_{i=1}^{m} f_i z_i + \sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij} y_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{q} c_{ijk} x_{ijk} \tag{1}$$

$$\sum_{i=1}^{m} x_{ijk} = 1, \forall j, \forall k \tag{2}$$

$$\sum_{i=1}^{m=1} z_i \le p \tag{3}$$

$$z_i - y_{ij} \le 0, \forall i, \forall j \tag{4}$$

$$-y_{ij}+x_{ijk}\le 0,\forall i,\forall j,\forall k \tag{5}$$

$$\sum_{j}\sum_{k}s_j d_{jk}x_{ijk}\le w_i,\forall i \tag{6}$$

$$\sum_{k}x_{i_r jk}=a_r,i_r\in\{1,2,\ldots,m\},\forall j \tag{7}$$

$$y_{ij}\ne 0,j=j_1,j_2,\ldots,j_r \tag{8}$$

$$x_{ijk}\ne 0,j=j_1,j_2,\ldots,j_r;k=k_1,k_2,\ldots,k_r \tag{9}$$

$$y_{ij}=\{0,1\},\forall i,\forall j \tag{10}$$

$$z_i=\{0,1\},\forall i \tag{11}$$

$$0\le x_{ijk}\le 1,\forall i,\forall j,\forall k \tag{12}$$

其中,下角标 i 是库存、销售机构等设施的集合，i=1,2,…,m；j 是产品(商品)的种类集合，j=1,2,…,n；k 是市场的顾客，也可以是销售网络终端，即末级销售点的集合，k=1,2,…,q 。供应链问题中的决策变量有 3 个，即表示可供使用的库存销售机构的变量，如果选中第 i 个设施，则 z_i=1,否则 z_i=0； y_{ij} 表示设施 i 关于产品 j 的配置变量，如果选中设施 i 对于产品 j 的配置，则 y_{ij}=1,否则 y_{ij}=0； x_{ijk} 表示设施 i 关于产品 j 对于市场顾客 k 的配置的份额，设施 i 于产品 j 对于市场顾客 k 配置是一种市场份额的表示，是一个在 0 到 1 之间的小数，实质上就是经过设施 i 处理的产品 j 在所有顾客市场占有率。供应链问题目标函数(1)中的参数 f_i 表示设施 i 的年度成本，g_{ij} 表示设施 i 关于产品 j 配置的年度成本，c_{ijk} 表示设施 i 关于产品 j 对于市场顾客 k 配置的年度成本。式(3)中的 p 是可供使用的设施的最大数量。式(6)中的 s_j 是产品 j 占用的空间; d_{jk} 是产品 j 对于市场顾客 k 单位时间的需求; w_i 是设施 i 的能力， w_i>0，它是来自于设施 i 可以提供总需求量的上限。式(7)中参数 $a_i\in[0,1]$是某一产品 j 市场占有份额。

在上述供应链问题中,目标函数(1)表示供应链管理过程中要使设施选址的成本、设施关于产品的配置的成本以及设施关于产品对于销售终端(市场顾客)配置的成本总和最小。条件式(2)表明每一个市场顾客的需求将得到满足。条件式(3)表明可供产品(商品)库存的设施数量不超过 p。约束条件(4)表明只有可供使用的设施 i 开放时，设施 i 才能对于产品 j 实行配置；否则，设施 i 不配置产品 j。条件式(5)表明只有设施 i 配置产品 j 时，市场顾客 k 才能从设施 i 得到产品 j；否则，市场顾客 k 不能从设施 i 得到产品 j。条件式(6)表明设施 i 的能力，即设施 i 总设计能力不超过 w_i。条件式(7)表明某一产品的分销占有市场的一定份额 a_i,,$0\leqslant a_i\leqslant 1$。条件式(8)表明必须保证某些产品对于设施 i 的配置，即，将产品 j 一定发送到设施 i。

条件式(9)表明必须保证某些产品对于市场顾客的配置，即将产品 j 通过设施 i 分送到市场顾客 k。

6.3 基于克隆算法的供应链优化算法

6.3.1 编码、约束条件、亲和度函数的处理

1.编码解码

供应链管理问题中有 3 个决策变量，其中选址变量 z_i，产品设施配置变量 y_{ij} 是(0,1)变量；产品设施市场顾客配置变量 x_{ijk} 是实数变量，$0 \le x_{ijk} \le 1$。因此。我们将 z_i、y_{ij} 直接处理为二进制串，其中 z_i 有 i 位，y_{ij} 有 i*j 位；我们将按具体值长度转化为二进制串，假设是最长为 x 位的小数，则使其串长为 l，l 满足

$$2^{l-1} < x_{ijk} * 10^x \le 2^l$$

则 x_{ijk} 可以表示为 m*n*q 个这样的二进制串。

解码时按照相反的顺序即可取出每个 z_i、y_{ij}、x_{ijk} 值。

2. 约束条件和亲和度函数的处理。

供应链管理问题中有多个约束条件。我们将约束条件转化为无约束优化问题处理，使所有约束条件变为平方项约束。比如约束条件（2）可变为

$$\Phi_2 = \alpha_2\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{q} x_{ijk} - 1\right)^2$$

在克隆算法操作中遇到不满足条件式(8)，式(9)的 y_{ij}，x_{ijk} 变量，则进行淘汰。约束条件式(10)～式(12)则在变量编码问题中自然解决了。这样我们可以得到供应链管理问题的无约束系统，其无约束函数,即

$$Z^a = Z + \sum_{i=2}^{7} \Phi_i$$

其中，Z 是供应链问题中的目标函数，Φ_i 是供应链问题中约束条件式(2)～(7)中相应转化的平方约束项。将无约束化的目标函数 Z^a 乘幂转化为亲和度

$$Z^s = \left(Z^a\right)^{-1}。$$

6.3.2 算法步骤

步骤一：种群初始化。

随机选取各个染色体个体。选取变异概率P_m^c，交叉概率P_σ^c，初始种群规模n，克隆规模N_c，迭代次数。

步骤二：对抗体种群进行克隆操作T_c^C。克隆过后，种群变为$\overline{A}'(k)=\{\overline{A_1}'(k),\overline{A_2}'(k),.....,\overline{A_n}'(k)\}$

步骤三：对目前种群$\overline{A}'(k)$进行克隆变异操作$\overline{A}''(k)=T_m^c(\overline{A}'(k))$，我们以概率$P_m^c$从$\overline{A}'(k)$中抽出抗体，对一个或多个属性进行实值变异,使其以一定概率随机变为其他属性值，删去此种群中不满足条件式(8)、式(9)的抗体。

步骤四：对目前种群$\overline{A}''(k)$进行克隆交叉操作$\overline{A}'''(k)=T_\sigma^c(\overline{A}''(k))$，交叉时我们使用离散重组法则，删去此种群中不满足条件式(8)、式(9)的抗体。

步骤五：对目前种群$\overline{A}'''(k)$进行克隆选择操作$\overline{A}(k+1)=T_s^c(\overline{A}'''(k))$。如果迭代次数满足停机条件，则停机并输出最优适应度函数及染色体解；否则，把此时种群作为下一代计算的初始抗体种群，转步骤二。

6.4 仿真试验与结果分析

我们以一个纯净水零售的供应链问题为例。假设批发商数目为 m=3，纯净水销售种类为 n=6，零售商数目为 q=15。供应链问题的目标函数(1)中的参数$(f_1,f_2,f_3)=[30,60,40]$(千元)，g_{ij}=[5,1,30,4,20,5,30,2,6,4,30,6,20,5,10,8,10,8](千元)，是一个 3*6 阶矩阵，这里写成行向量，以便于计算机读入。这里的 3 行 6 列矩阵，每 6 个元素相当于 1 行，一共 3 行。下面的矩阵读入方式也是这样的。约束条件(3)中 p=6，约束条件(6)中s_j=[6,3,5,3,4,4]，d_{jk}=[0.12,0.43,0.5,……]（千件）(商品)，其总共有 6*15 个值，w_i=[14,17,16](千件)，约束条件(7)中，$y_{ij}\neq 0, j=1$，约束条件(8)中，$x_{ijk}\neq 0, j=1, k=1,2,...,15$，$c_{ijk}=\{[0.12,0.2,...]...\}_{3\times6\times15}$(千元)。

我们可以取变异概率P_m^c为 0.2，交叉概率P_σ^c为 0.5，初始种群 n 选为 200 个，克隆规模N_c选为 500 个，停机条件为当连续七十代适应度值不再变化。则我们在第 243 代可以得到最优解值，解码后可得到以下最优解：

$(z_1,z_2,z_3)=(1,0,1), y_{ij}=(0,1,0,0,1,0,1,0,0,0,0,0,1,1,0,0,1,1)_{3\times6}$,

$x_{ijk}=\{[0.323,0.627,...]..\}_{3\times6\times15}$,

其对应的最优解为$Z'=3.2537e-3$。

而以同样参数执行的遗传算法在第 243 代则得到以下解：

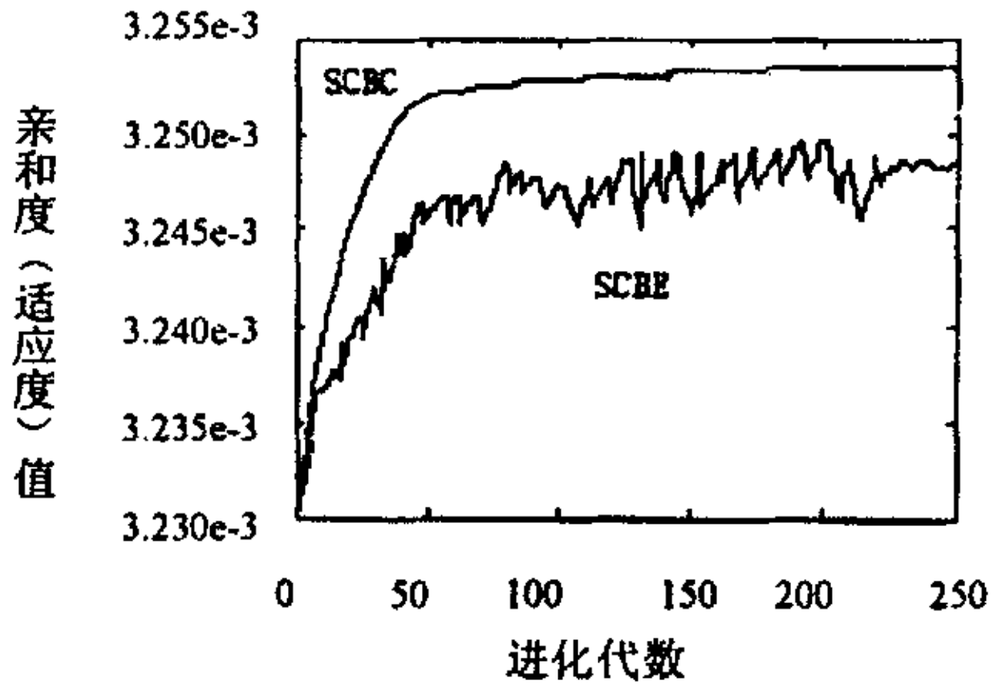
$(z_1,z_2,z_3)=(1,0,1), y_{ij}=(1,1,0,0,1,0,1,0,0,1,0,0,1,0,0,0,1,0)_{3\times6}$,

$x_{ijk}=\{[0.4,0.713,...]..\}_{3\times6\times15}$,

其对应的解为$Z'=3.2477e-3$

由图 6.1 可以看出，基于多克隆选择的供应链求解算法比用遗传算法求解的供应链算法的执行性能要好，其最佳亲和度相比于遗传算法的最佳适应度具有明

显的优势。而且，其收敛速度要快于基于遗传算法的供应链求解算法，在进化代数相同的情况下，其更快的收敛到最优解。



SCBC 基于克隆算法的供应链求解算法
SCBE 基与遗传算法的供应链求解算法

图 6.1 两种供应链求解算法最佳亲和度（适应度）比较

6.5 结论

通过实验我们可以发现基于多克隆选择算法的供应链求解算法具有收敛速度快的特点，这样我们就可以适当减少算法的迭代次数，从而提高供应链求解效率。另外，由于多克隆算法本身具有较强的全局及局部搜索能力、良好的并行性，因此将多克隆选择算法用于供应链问题的求解是有意义的。

第七章 结 论

需求是发明之母。

随着计算机硬件和数据库技术的发展，从海量的数据中挖掘出人们需要的知识变为可能，担负这一使命的数据挖掘技术便应运而生。数据挖掘技术主要融合了人工智能、统计学、数据库、机器学习等领域的技术，可以说，有数据的地方就需要数据挖掘。因此，数据挖掘技术引起了学术界和产业界的极大关注，吸引着越来越多的研究人员参与到这一领域技术的研究和开发。

现代企业在发展过程中，越来越意识到物流领域已成为整个行业发展的瓶颈。由于物流领域的技术含量低于社会平均水平，因此它严重阻碍了生产与流通领域的顺畅发展。供应链及其管理正是针对这一现状而产生的，由于它可以带来巨大的经济效益，各种各样的新技术被应用了进来。而这些算法只是关心供应链如何建模，至于如何求解供应链模型，大多数人们仍然采用单一的进化算法来求解，如何提高供应链模型的求解效率日益成为需要深入研究的课题。

本文所作的主要工作：

（1）为了克服 Apriori 算法要产生大量的候选项集，且每次都需要重复的扫描数据库的问题及基于进化算法的关联规则挖掘算法易陷入早熟、局部搜索能力差的问题，我们应用免疫克隆算法对多维关联规则挖掘进行分析得到聚类新算法，针对遗传算法收敛速度过慢这一点，我们引进克隆算子提出了将免疫克隆算法应用到多维关联规则挖掘中来，文中给出了改进后的算法，它是在保留遗传算法优良特性的基础上有目的、有选择的利用待求问题中的一些特征信息或先验知识来抑制进化过程当中出现的种群退化现象。实验证明，它能够使多维关联规则挖掘结果收敛到全局最优，同时相比 Apriori 算法和标准遗传算法，具有更快的收敛速度和更强的搜索能力。

（2）为了克服基于进化算法的供应链求解算法收敛速度慢、易早熟的问题，我们提出了利用免疫克隆算法来处理聚类问题的一种新算法 SCBC。克隆选择算法兼顾全局搜索与局部搜索，有更好的种群多样性，具有记忆性能，能保证以概率 1 收敛到最优解，并且它弥补了进化算法较少关注种群间的协作这一缺点，因此将它用于供应链求解问题是有理论意义的。实验证明，基于克隆算法的供应链求解算法收敛速度快、搜索能力强，可以大大提高供应链模型求解的效率。

综上所述，本文对数据挖掘及其关联规则挖掘技术、供应链求解问题进行了系统地分析和深入研究，获得了一系列有价值的研究成果。在此基础上，作者认为下一步要努力的方向为：

对目前的关联规则挖掘而言，它还存在以下问题：在处理极大量的数据时，

如何提高算法效率的问题；对于挖掘迅速更新的数据的挖掘算法的进一步研究；在挖掘的过程中，提供一种与用户进行交互的方法，将用户的领域知识结合在其中；对于数值型字段在关联规则中的处理问题；生成结果的可视化方面等等。

对供应链管理及其模型求解而言，快速响应顾客需要、进一步提高供应链模型的求解效率等问题依然突出，而供应链模型的系统性和灵活性依然是人们关心的话题，因此研究适合我国国情并具有快速反应能力及快速求解能力的供应链依然是科技研究人员以后研究的主要方向。

最后不得不提的是，由于本人水平和学识有限，研究中的不足和问题在所难免，恳请各位老师批评指正。

致 谢

本论文是我攻读硕士学位期间所做工作的总结。在这里，谨向所有辅导和帮助过我的老师以及同学们致以诚挚的谢意。

感谢我的导师刘芳教授，她对我的指导使我获益很多。在论文的选题及研究过程中也得到了刘老师耐心的指导。她对工作的认真态度和对学生负责的态度也深深地影响着我。

感谢我的同学。在和大家一起生活、一起工作的两年多时间里，彼此之间取长补短、增长知识，和大家的交流让我更加体会到同学之间真挚的友谊、体会到团结的力量有多么强大。

最后也感谢我的父母，父母的支持会使我更加坚强。

感谢所有帮助过我的人们，祝他们生活得越来越美好！

参考文献

[1] （加）Jiawei Han, Michelle Kamber《数据挖掘概念与技术》，机械工业出版社，2001。

[2] Agrawal R., Imiclinski T., Swami A. Database Mining:A performance perspective. IEEE Trans. Knowledge and Data Enginnering.1993,5.914~925.

[3] Agrawal R., Srikant R. Fast Algorithm for Mining Association Rules.In Proceeding 1994 International conference Very Large Data Bases(VLDB'94). Santiago, Chile.Sept,1994.487~499.

[4] Mannila H.,Toivonen H.,Verkamo A.I. Efficient Algorithm for Discovering Association Rules. In Proceedings AAAI'94 Workshop Knowledge Discovery in Databases(KDD'94). Seattle WA.1994.181~192.

[5] Agrawal R., Shafer J.C. Parallel Mining of Association Rules: Design, Implementation, and Experience. IEEE Trans. Knowledge and Data Engineering. 1996,8. 962~969.

[6] Toivonen H. Sampling Large Databases for Association Rules.In Proceeding 1996 International conference Very Large Data Bases(VLDB'96). Bombay, India. Sept.1996.134~145.

[7] Brin S., Motwani R., Ullman J.D. et al. Dynamic Itemset Counting and Implication Rules for Market Basket Analysis. In Proceedings of ACM-SIGMOD International Conference Management of Data(SIGMOD'97).Tucson, AZ. 1997.255~264.

[8] Han Eui-Hong, George K., Kumar V. Scalable Parallel Data Mining for Association Rules. Proceeding of the ACM SIGMOD97,1997. 277~288

[9] Agrawal R., Srikant R. Mining Sequential Patterns. In Proceedings International Conference Data Engineering (ICDE'95).Taibei.Taiwan.1995.3~14.

[10] Koperski K.,Han J. Discovery of Spatial Association Rules in Geographic Information Databases.In Proceedings 4th International Sympothsis Large Spatial Databases(SSD'95). Portland,ME.1995.47~66.

[11] Lu H.,Han J.,Feng L. Stock Movement and n-dimentional Inter-transaction Associaltion Rules. In Proceedings SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98). Seattle,WA.1998.12:1~12:7.

[12] Bayardo R.J. Efficiently Mining Long Patterns from Databases.In Proceedings ACM-SIGMOD International Conference Management of Data(SIGMOD'98). Seattle,WA.1998.85~93.

[13] Pasquier N.,Bastide Y.,Taouil R. et al. Discovering Frequent Closed Itemsets for Association Rules.In Proceedings 7th International Conference Databases Theory(ICDT'99). Jerusalem,Israel. 1999.398~416.

[14] Beyer K.,Ramakrishnan R. Bottom-up Computation of Sparse and Iceberg Cubes. In Proceedings ACM-SIGMOD International Conference Management of Data (SIGMOD'99). Philadelphia, PA. 1999.359~370.

[15] Han J.,Pei J.,Yin Y. Mining Frequent Patterns without Candidate Generation.In Proceedings ACM-SIGMOD International Conference Management of Data (SIGMOD'00). Dallas,TX. 2000.1~12.

[16] 李群明，张士廉，王成恩，基于 COBRA 的多 Agent 供应链管理系统的研究，

信息与控制，2000 年第 6 期。
[17] 王春喜，查建中，李建勇，鄂明成，面向网络制造的知识供应链建模，计算机集成制造系统，2000 年第 6 期。
[18] 赵耀华，以核心业务为中心的敏捷供应链研究，淮海工学院学报，2000 年第 4 期。
[19] 袁 磊，陈长青，冯玉才，数据仓库的信息供应链模型，计算机工程与应用，2001 年 22 期。
[20] 曹文彬，何建敏，流程企业供应链管理的建模问题研究，制造业自动化，2001 年第 9 期。
[21] 邵晓峰，季建华，黄培清，面向大规模定制的供应链模型的研究，制造业自动化，2001 年第 6 期。
[22] 韩向东，张春远，U.W.Geitner，基于标准过程的供应链管理模型的研究，江苏理工大学学报，2000 年第 3 期。
[23] 卢 震，黄小原，栗东生，一类供应销售条件下的供应链模型与决策应用，系统工程理论方法应用，2000 年第 4 期。
[24] 黄小原，卢 震，一种交互式进化规划在供应链方向的应用，东北大学学报，2000 年第 5 期。
[25] 曹 杰，王红卫，供应链联合优化数学模型及求解的混合算法，决策借鉴，2001 年第 5 期。
[26] 卢 震，黄小原，服务销售系统供应链模型设计与应用，东北大学学报，2001 年第 6 期。
[27] Dasgupta D. Artificial immune systems and their applications [M] . Berlin: Springer Verlag , 1999. 3~23.
[28] De Castro , Von Zuben. Artificial immune systems ， 1999 12.
[29] 王 磊, 潘 进, 焦李成. 免疫算法[J]. 电子学报, 2000, 28(7)： 74-78.
[30] 王 磊, 潘 进, 焦李成. 免疫规划[J]. 计算机学报, 2000, 23(8)： 806-812.
[31] 徐 科. 神经生物学纲要[M]. 北京： 科学出版社, 2000. 374-386.
[32] 于善谦, 王洪海, 朱乃硕, 等. 免疫学导论[M]. 北京:高等教育出版社, 1999. 5-9.
[33] Ishida Y. International workshop on the immunity based systems 1996 (IMBS' 96) held in conjunction with ICMAS' 1997 5-27.
[34] Farmer J D , Packard N H , Perelson A S. The immune system, adaptation , and machine learning [J] . *Physica D* ,1986 ,22 :187-204.
[35] 杜海峰、焦李成，人工免疫系统中的克隆选择算法，西安电子科技大学智能信息处理研究所；2003.11。
[36] P.J.Bentley and J.P.Wakefield. Overview of Genetic Evolutionary Design Systems, In Proceedings of the 2nd On-Line World Conference of Evolutionary Computation (WEC2),pp. 53-56, 1996.
[37] 陈国良，王煦法，庄镇泉等. 遗传算法及其应用。北京：人民邮电出版社，1996.
[38] 林学颜， 张玲. 现代细胞与分子免疫学. 北京：科学出版社，1999.
[39] 陆德源，马宝骊. 现代免疫学. 上海：上海科技教育出版社，1998.
[40] 杜海峰, 王孙安. 基于 ART—人工免疫网络的数据浓缩方法研究 . 模式识别与人工智能, 2001,14(4): 401~405.

[41] Hunt J E , Cooke D E. Learning using an artificial immune system[J] . *Journal of Network and Computer Applications* , 1996 ,19 (2) :189-212.
[42] Timmis J , Neal M, Hunt J . Data analysis using artificial immune systems , cluster analysis and kohonen networks systems , cluster analysis and kohonen networks : some comparisons [A] . In : *IEEE SMC ' 99 ConferenceProceedings*[C] . Piscataway , NJ , USA: IEEE ,1999. 922-927.
[43] Dasgupta D. Immunity2based intrusion detection system: a genera framework [A] . In : *Proceedings of 22 nd National Information System Security Conference* [C] . Gaithersburg ,USA: NIST,1999. 147 160.
[44] 张 军, 刘克胜, 王煦法. 一种基于免疫调节和共生进化的神经网络优化设计方法[J] . 计算机研究与发展,2000 , 37(8) : 924 930.
[45] T. M. Mitchell: Machine Learning and Data Minging, Communications of the ACM ,Vol .42, No .11, 30-36,November 1999.
[46] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Conference on Management of data, pp. 207-216, 1993.
[47]R.Agrawal, and J. Shafer. Parallel mining of association rules: Design, Implementation, and Experience. Technical Report FJ10004, IBM Almaden Research Center, San Jose, CA 95120, Jan. 1996.
[48]S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets:generlizing association rules to correlations. Proceedings of the ACM SIGMOD, 1996. pages 255-276.
[49] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset counting and implication rules for market basket data. In ACM SIGMOD International Conference On the Management of Data. 1997.
[50] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos. Ratio rules: A new paradigm for fast, quantifiable data mining.
[51] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Segmentation problems. Proceedings of the 30th Annual Symposium on Theory of Computing, ACM. 1998.
[52] J. L. Lin, and M. H. Dunham. Mining association rules: Anti-skew algorithms. Proceedings of the International Conference on Data Engingeering, Orlando, Florida, February 1998.
[53] H. Mannila, H. Toivonen, and A. Verkamo. Efficient algorithm for discovering association rules. AAAI Workshop on Knowledge Discovery in Databases, 1994, pp. 181-192.
[54] J. S. Park, M. S. Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. Proceedings of ACM SIGMOD International Conference on Management of Data, pages 175-186, San Jose, CA, May 1995.
[55] J. S. Park, M. S. Chen, and P. S. Yu. Efficient parallel data mining of association rules. 4th International Conference on Information and Knowledge Management, Baltimore, Maryland, Novermber 1995.
[56] R. Srikant, and R. Agrawal. Mining generalized association rules. Proceedings of the 21st International Conference on Very Large Database, 1995, pp. 407-419.
[57] R. Srikant, and R. Agrawal. Mining quantitative association rules in large relational

tables. Proceedings of the ACM SIGMOD Conference on Management of Data, 1996. pp.1-12.
[58] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very large Database, 1995.
[59] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of costomer transactions. Proceedings of the International Conference on Data Engineering, February 1998.
[60] H. Toivonen. Sampling large databases for association rules. Proceedings of the 22nd International Conference on Very Large Database, Bombay, India, September 1996.
[61] M. J. Zaki, S. Parthasarathy, and W. Li. A localized algorithm for parallel association mining. 9th Annual ACM Symposium on Parallel Algorithms and Architectures, Newport, Rhode Island, June 1997.
[62] J.Han, J.Pei, and Y.Yin. Mining frequent patterns without candidate generation. In Proc.2000 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD'00), Dalas, TX, May 2000.
[63] Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D.Ullman, Cheng Yang. Finding Interesting Associations without Support Pruning.
[64] Jiawei Han, Sonny H.S. Chee, Jenny Y.Chiang. Issues for On-Line Analytical Mining of Data Warehouses.
[65] Information Discovery, Inc. OLAP and DataMining, Bridging the Gap.
[66] C. C. Aggarwal, and P. S. Yu. A new framework for itemset generation. IBM Research Report,RC-21064.
[67] Kim, J., Ong, A., and Overill, R., Design of an Artificial Immune System as a Novel Anomaly Detector for Combating Financial Fraud in Retail Sector, Proceeding of the Congress on Evolutionary Computation (CEC-2003), Canberra:2003,405-412.
[68] 宋华，胡左浩，现代物流与供应链管理，经济管理出版社，2000年4月。
[69] Thomas Douglas J,Griffin Paul M. Coordinated supply chain management [J].European Journal of Operations Research,1996,94:1～15.
[70] Rogers D F,Tsubakitani S.Newsboy-style results for multi-echelon inventory problems:Back-orders optimization with intermediate delays[J].Journal of the OR Society,1991,42(1):57～68.
[71] Lee H,Rosenblatt M J.A generalized quantity discount pricing model to increase supplier profits[J].Management Science,1986,32(9):1177～1185.
[72] Anupindi R, Akella R. Diversification under supply uncertainty[J].Management Science,1993,39(8):944～963.
[73] Banerjee A.A joint economic lot size model for purchaser and vendor [J].Decision Science,1986,17:292～311.
[74] Goyal S K.A joint economic lot size model for purchaser and vendor : acomment [J].Decision Science,1988,19:236～241.
[75] Kohli R,Park H.Coordination Buyer-seller transactions across multiple products[J].Mangement Science,1994,40(9):45～50.
[76] Lau H,Lau A H-L.Coordinating two suppliers with offsetting lead time and price

performance[J].Journal of Operations Management,1994,11:327～337.
[77] Chien T W.Determining profit-maximizing production/shipping policies in a one-to-one direct shipping,stochastic environment[J].European Journal of Operations Research,1993,64:83～102.
[78] Chandra P,Fisher M L.Coordinating of production and distribution planning[J].European Journal of Operations Research,1993,72:503～517
[79] Clark A J,Scarf H.Optimal policies for a multi-echelon inventory problem[J].Management Science,1960,6:475～490.
[80] Silver E,Peterson R.Decision systems for inventory management and production planning[M].NewYork:Wiley,1985.
[81] Muckstadt J A,Roundy R O.Logistic of production and inventory [M]. North-Holland,Amsterdam:1993.
[82] Lee H,Billington C.Managing supply chain inventory:pitfalls and opportunities [M] .Sloan Management Review(Spring1992):65~73.
[83] 方述诚,普森普拉著.汪定伟,王梦光译.线性规划及扩展[M].理论与算法.北京:科学出版社.1994.

作者读研期间论文成果

[1] 刘芳，孙杨军. 基于多克隆选择的多维关联规则挖掘算法，复旦大学学报，vol43,no.5,2004.10, p742-745.